PROBABILITES ET STATISTIQUES
STATISTIQUES INFERENTIELLES
(BTS)

B. Bigot
B. Chaput
J-C. Daniel
J-C. Duperret

Université de
REIMS

Institut de
Recherche sur
l'Enseignement
des
Mathématiques

Moulin de la Housse BP 1039
51687 REIMS CEDEX 2
Tél : 03 26 05 32 08 - Fax : 03 26 85 35 04

PROBABILITES ET STATISTIQUES

STATISTIQUES INFERENTIELLES

(B.T.S.)

Bernard BIGOT

Brigitte CHAPUT

Jean-Claude DUPERRET

Jean-Claude DANIEL

I.R.E.M. de REIMS

SOMMAIRE

INTRODUCTION

Les statistiques et, à un degré moindre, les probabilités ont souvent été les parents pauvres de l'enseignement traditionnel des mathématiques. Un certain mépris des "puristes" les a souvent reléguées au rang d'un enseignement de recettes. Et pourtant, elles font partie de ce que les spécialistes appellent mathématiques "appliquées" qui, paradoxalement, font appel à des théories souvent bien compliquées. Et pourtant, elles sont un formidable outil d'"intelligibilité" du monde ; Laplace, en 1812, l'affirmait déjà :

> *"Et si l'on observe ensuite que dans les choses qui peuvent ou non être soumises au calcul, la théorie des probabilités... apprend à se garantir des illusions. Il n'est pas de science qu'il soit plus utile de faire entrer dans le système de l'instruction publique."*

Souvent délaissées dans les cursus scientifiques traditionnels, les statistiques et les probabilités sont au contraire le noyau dur et, souvent, la seule présence des mathématiques dans de nombreuses formations post-bac. Mais confiées à des enseignants non nécessairement formés, disposant en général d'un temps d'enseignement réduit, s'adressant à un public non scientifique parfois en échec mathématique, elles ont gardé cet aspect "recettes".

Cela a été le cas dans beaucoup de sections de Techniciens Supérieurs où les enseignants se sont dans un premier temps auto-formés très rapidement pour assurer cet enseignement qu'on leur confiait. Ce sont ces mêmes enseignants qui ont compris que, si l'on voulait dépasser un enseignement algorithmique, il importait de donner du sens aux concepts que l'on construisait. Ce sont ces enseignants qui ont souhaité une formation qui les aide à la fois à mieux maîtriser le modèle théorique mais aussi la modélisation, enjeu profond du passage des statistiques aux probabilités.

Ce besoin exprimé ne pouvait en aucun cas être satisfait par un stage paraphrasant les programmes officiels et proposant quelques exercices et sujets de BTS. C'est d'abord à un travail de

recherche que doit se livrer l'équipe de futurs formateurs. Au sein de l'IREM de Reims, sollicitée pour cette formation, notre équipe s'est constituée pour cette tâche.

Notre premier travail a été un travail de lecture. La bibliographie de cette brochure donne la liste des ouvrages sur lesquels s'est appuyée notre réflexion. Elle n'est donc pas exhaustive. Au-delà des manuels scolaires ou universitaires classiques, nous voulons mettre en avant quelques documents que nous avons particulièrement utilisés.

- Tout d'abord le travail des IREM, en particulier Besançon et la commission Inter-IREM Lycées Techniques, ainsi que celui de l'APMEP.
- Des dossiers pour la formation continue, de Michel CHAVIGNY, édités au CAFOC de Besançon, en particulier pour l'estimation où l'on retrouve la plupart des situations proposées.
- Le livre de Michel LAVIEVILLE, dont l'originalité pour les "matheux" est qu'il s'adresse à des futurs médecins, en particulier pour le test du Khi-Deux.
- Le magnifique livre d'Arthur ENGEL, qui mêle intimement problèmes, conjectures, simulations, compréhension de phénomènes "naturels" et résultats mathématiques.
- Le très beau livre d'Ivar EKELAND, pour le sens profond qu'il donne.

Au risque d'être accusés de plagiat, nous n'avons pas voulu modifier les problèmes et les situations que nous avons exploités lorsqu'ils nous paraissaient pertinents par rapport à notre propos. Notre souci commercial étant nul, nous avons au contraire trouvé cela plus honnête intellectuellement, en souhaitant inciter le lecteur à se procurer ces ouvrages pour les découvrir plus complètement.

Si notre second travail a été la préparation et l'animation elle-même des stages, notre dernière tâche (qui ne fut pas la moindre !), c'est cette brochure. Passer des supports prévus pour un exposé oral à une production écrite et autonome oblige à tout repenser, mais nous y avons été poussés très fortement par nos stagiaires. C'est du reste la grande richesse des IREM que de pouvoir assurer, chaque fois que cela est possible, cette trilogie : recherche-formation-production. Il nous est apparu important de proposer à la fois des activités directement utilisables en classe par les étudiants et la réflexion qui s'y rattache, plus tournée vers les enseignants.

Dans une première partie "Partons au hasard", nous posons le problème fondamental, celui de la modélisation, c'est-à-dire le passage de l'observation par les statistiques au traitement dans le modèle mathématique des probabilités.

La seconde partie illustre cet aller-retour : statistiques-probabilités-statistiques en rappelant les lois de probabilité usuelles et en donnant à travers de nombreux exemples leur cadre et leurs limites d'utilisation. Cette partie débouche sur l'échantillonnage qui ouvre la voie aux statistiques inférentielles.

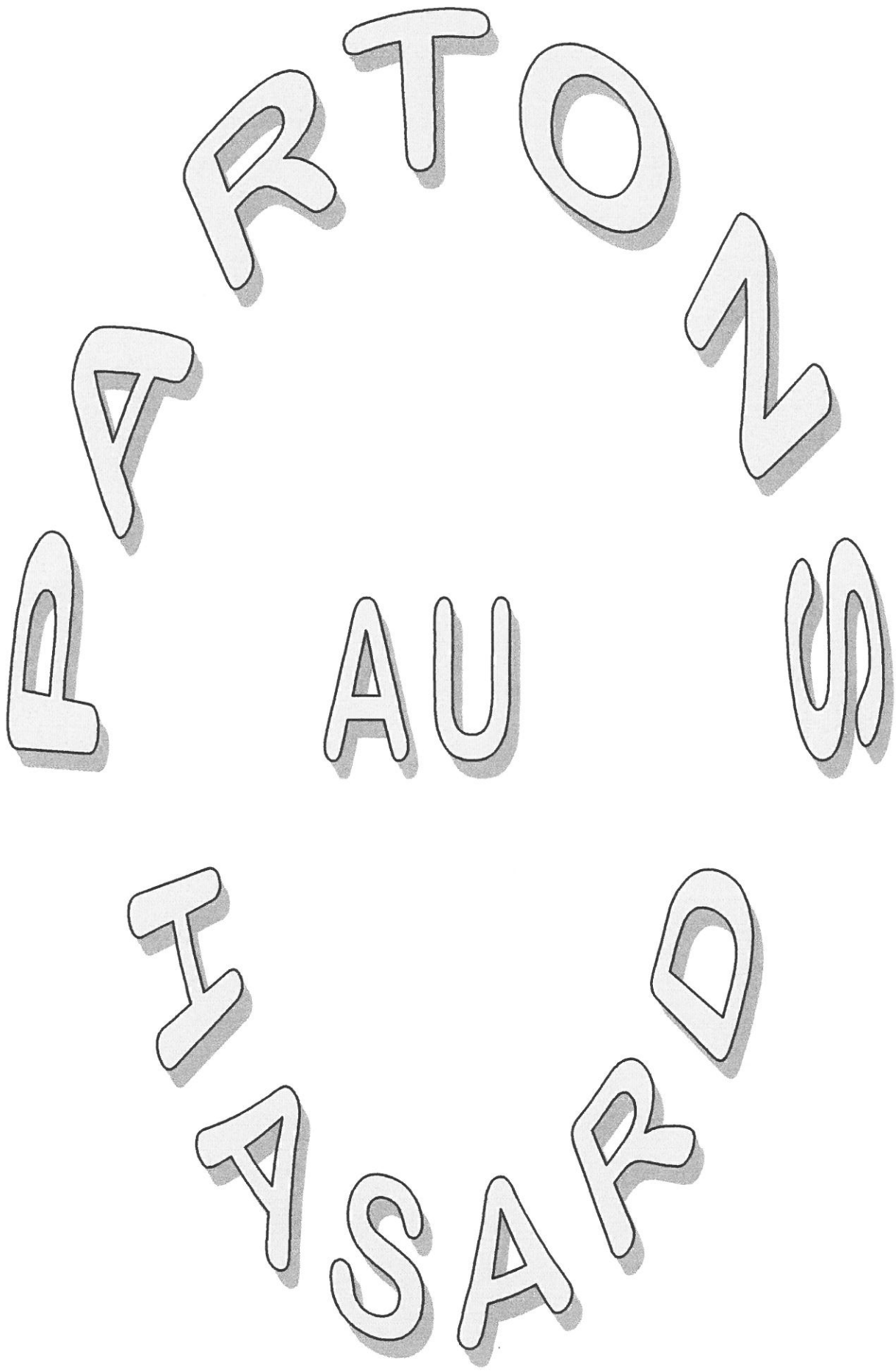
Ces statistiques inférentielles sont l'objet de la troisième partie : comment à partir d'un échantillon obtenir des renseignements sur la population totale (estimation) ? avec quel degré de confiance (tests) ? Nous y abordons alors un problème fondamental de l'économie actuelle, celui de la fiabilité ou, en terme plus mathématique, la probabilité de panne d'un dispositif.

Quelques réflexions sur des sujets d'examens proposés en BTS nous ramènent aux problèmes de l'enseignement et de l'évaluation. Et pour terminer, une partie, essentiellement de référence, regroupe des tables de lois de probabilité, le programme de BTS et un formulaire.

Bernard BIGOT, Brigitte CHAPUT,
Jean-Claude DANIEL, Jean-Claude DUPERRET

C'est certainement ici l'endroit pour remercier Pascal GRISONI et Michel HENRY qui ont bien voulu assurer relecture et corrections de cette brochure.

Cette brochure a été réalisée en partie grâce au soutien de la Direction des Lycées et Collèges dans le cadre des contrats DLC ADIREM (thème BTS).


PARTONS
AU
HASARD

Situons le problème

Introduction

On appelle statistique descriptive l'étude et l'élaboration de données recueillies au cours d'une série particulière d'observations (échantillon).

La statistique doit permettre de passer de cet ensemble d'observations à l'ensemble plus général (population) dont il est issu.

La statistique est donc l'ensemble des méthodes utilisées pour obtenir des renseignements sur une population à partir de renseignements sur un échantillon de cette population. L'échantillon sera supposé « tiré au hasard » et cela méritera d'être précisé.

Il faut éviter de confondre « techniques statistiques » en tant que science, branche des mathématiques, avec les « statistiques » en tant données relatives à des dénombrements ou inventaires.

Statistique et probabilité

Les statistiques reposent dans leurs fondements sur la théorie des probabilités. Nous verrons cependant que les probabilités elles-mêmes peuvent s'introduire de deux façons : a priori, avec toute la place laissée à la subjectivité, ou par fait d'expériences répétées (aspect « fréquentiste »). Les statistiques se réfèrent à la théorie des probabilités et les probabilités se nourrissent des statistiques. Cette dialectique les conforte l'une l'autre plus qu'elle ne les oppose.

La notion de variable aléatoire sera au coeur de la modélisation de phénomènes où intervient le hasard.

L'indépendance formalise le fait que les tirages successifs s'effectuent dans certaines conditions. C'est une notion difficile et source de bien des erreurs chez les étudiants de STS.

Le passage de renseignements sur un échantillon à des renseignements sur la population et réciproquement, sera abordé dans le contexte de l'estimation et des tests d'hypothèses.

Echantillon

Quelle que soit la puissance des techniques statistiques, elles ne permettront d'arriver à quelque conclusion que ce soit, qu'à partir d'informations qu'il faut aller chercher :

- où et comment prélever des échantillons de données ?
- comment extrapoler à tout un groupe, l'observation faite sur un nombre limité d'individus ?

Le problème est évidemment moins banal qu'il n'y pourrait paraître, comme en témoigne l'exemple suivant, cité par Georges Parreins dans « Techniques Statistiques » :

« Comptant prouver la nécessité d'avoir un médecin au moment d'un accouchement, on questionne des femmes ayant donné naissance à un enfant. Pour cinquante accouchements avec médecin on constate quatre complications et pour cinquante accouchements sans praticien, on ne constate que trois complications.

Ce résultat soumis à un statisticien engendre la réponse immédiate : les résultats ne sont pas significatifs et pour tirer une conclusion sérieuse, il faudrait opérer sur des effectifs plus importants.

Un nouveau recueil d'observations est alors effectué. Sur deux séries de cinq cents accouchements, on trouve 47 complications dans le premier groupe (avec médecin) et 19 dans l'autre.

Le statisticien consulté à nouveau déclare que dans ce cas les résultats sont significatifs et qu'avec un risque très faible de se tromper (de l'ordre de 1 sur 10 000), il y a moins de complications en l'absence de médecin.

Que faut-il en penser, et que penseraient les étudiants de STS ?

La conclusion peut paraître pour le moins inquiétante. Il faut évidemment chercher l'explication dans le « comment » de l'expérimentation. Cerner la population dans laquelle on choisit l'échantillon représente une phase essentielle. En fait ici, les deux échantillons avaient été prélevés dans une population rurale où l'habitude culturelle est de n'appeler le médecin qu'en cas de risque de complications !

Le prélèvement d'échantillon ne s'est donc pas effectué « au hasard ». Mais si le résultat avait joué en sens inverse, qui aurait demandé un complément d'enquête ?

Il faut donc redire avec Claude Bernard, que l'expérimentateur doit toujours douter, fuir les idées fixes et garder sa liberté d'esprit.

Les échantillons non significatifs, mal prélevés, sont la cause d'un grand nombre de conclusions erronées et ce, dans beaucoup de cas qui ne sont pas des cas d'école.

Interprétation

Les informations ayant été correctement prélevées, comment les exploiter ? En fait que cherche-t-on ?

On cherche finalement le plus de renseignements possibles qui peuvent se regrouper en trois rubriques :

1. Etudier les liaisons entre diverses grandeurs

Les lois de la nature (scientifiques, économiques, humaines...) sont souvent difficiles à comprendre ou à interpréter. On cherche souvent si un phénomène est influencé par certains facteurs dont on soupçonne l'existence et si oui, quelle est la forme de la liaison.

En se limitant au cas de deux grandeurs X et Y, deux cas peuvent se présenter :

a. X et Y sont toutes deux aléatoires.

On effectue une étude de corrélation qui fournit seulement une indication sur les variations simultanées des deux variables. Deux variables indépendantes ont une corrélation nulle, la réciproque est fausse !

b. L'une des variables seulement est aléatoire, l'autre lui est liée.

On effectue une étude de régression bien adaptée au principe de causalité. Il faut cependant se souvenir que la causalité ne se prouve pas par une formule mathématique. C'est un fait d'expérience, donc lié à l'observation.

2. Comparer entre eux deux ou plusieurs paramètres.

Comparer deux paramètres, deux grandeurs, deux qualités, c'est décider s'il y a ou non égalité, ou faire un choix entre deux états. La théorie des tests permet dans certaines conditions de porter de tels jugements et donc d'aider à la prise de décision. De tels tests et prises de décisions comportent toujours des risques :

- risque de première espèce qui est celui de refuser un résultat en fait correct,
- risque de deuxième espèce qui est celui d'accepter un résultat en fait erroné.

Cette théorie des tests prend évidemment beaucoup d'importance dans de nombreux domaines et en particulier dans les processus de contrôle industriel des fabrications.

3. Estimer certaines grandeurs.

On est toujours conduit à énoncer l'estimation sous la forme : « il y a k chances sur 100 pour que la valeur de la grandeur considérée soit comprise entre a et b ». [a,b] est un intervalle de confiance au seuil indiqué. Il est à remarquer que la largeur de l'intervalle [a,b] varie avec la taille de l'échantillon pour un seuil donné.

Le risque

En statistiques, plus qu'ailleurs, il faudra aider les étudiants à se garder de trop de certitudes. La théorie de la décision ne requiert nullement que l'on ait des bases objectives pour les probabilités.

Ivar Ekland (« Au hasard », Seuil) donne l'exemple suivant qui éclaire bien le propos :

« Je peux être parfaitement convaincu que la fin du monde interviendra demain. J'attribuerai à cet événement la probabilité 1 et je serai conduit à agir en conséquence. »

Il s'agit là d'une probabilité subjective, mais ce n'est pas parce qu'une conviction est irrationnelle qu'elle a moins de force et qu'elle n'entraîne pas de prises de décisions.

Nous pouvons donc être confrontés à deux types de situations :

- une situation où l'on dispose d'un modèle probabiliste exact; ce sont ces situations que nous qualifierons d'aléatoires,

- une situation d'ignorance ou de méconnaissance. Par exemple on demande à une personne de parier sur le résultat d'un match de tennis entre joueurs de forces inégales quand elle ne connaît rien sur le classement de ces deux joueurs.

L'expérience d'Ellberg relatée par I. Ekland met bien en évidence ce risque d'ignorance dû à une absence d'information :

> *L'expérience d'Ellsberg est la suivante. On présente aux sujets deux urnes, chacune contenant 100 boules. On annonce que la première contient exactement 50 boules rouges et 50 boules noires; quant à la seconde, on dit simplement qu'elle ne contient que des boules noires et/ou des boules rouges, sans préciser dans quelle proportion.*
>
> *On ouvre alors les paris sur la première urne. Les sujets choisissent une couleur puis l'on tire une boule. Ceux qui ont deviné juste gagnent 100$, les autres rien. L'expérience prouve que la plupart des gens parient indifféremment sur le noir ou sur le rouge, c'est-à dire que leurs probabilités subjectives sont 0,5 et 0,5.*
>
> *Vient ensuite une série de paris sur la deuxième urne. Les conditions du pari sont les mêmes, mais cette fois les sujets n'ont aucun renseignement sur le contenu de l'urne hormis le fait que la boule qui sortira sera noire ou rouge. C'est donc une situation d'ignorance, par rapport à la précédente qui était une situation aléatoire. Conformément à la théorie, la plupart des gens parient indifféremment sur le rouge ou sur le noir, c'est-à-dire qu'ils leur attribuent encore les probabilité de 0,5 et 0,5.*
>
> *Arrive enfin le paradoxe. On ouvre une troisième séance de paris. On gagne 100$ pour une boule rouge et rien pour une boule noire, mais on a le droit de choisir l'urne d'où la boule sera tirée. Comme les probabilités subjectives sont les mêmes dans les deux cas, et que les sujets estiment donc que dans un cas comme dans l'autre ils ont une chance sur deux de gagner, la théorie exigerait que les sujets choisissent indifféremment l'une ou l'autre des deux urnes. Or il n'en est rien, la plupart des gens expriment une préférence marquée pour la première urne (celle dont les proportions sont connues). Cette préférence est encore plus marquée s'il s'agit de perdre 100$ plutôt que de les gagner ! - mais on trouve moins de volontaires. Tout se passe donc comme si l'ignorance était un facteur de risque supplémentaire que les probabilités subjectives à elles seules n'arrivent pas à prendre en compte.*

Indépendance.

Au coeur de toute analyse statistique, on trouve la notion d'indépendance, liée au temps, à la mémoire, à l'espace, aux « systèmes ». Notion bien délicate, et source de biens des difficultés.

Ainsi, s'il pleut aujourd'hui sur Reims en même temps qu'il pleut sur Paris, quel sentiment avons-nous sur la dépendance de ces deux événements ? Mais s'il pleut aujourd'hui sur Reims en même temps qu'il pleut sur Tokyo notre sentiment sera-t-il le même? Serait-t-il le même pour un astronaute en orbite autour de la terre?

Le statisticien sera donc toujours confronté à un problème d'interprétation quand il isole des événements dont il postule qu'ils sont aléatoires. Il devra confronter son modèle à la réalité. Son travail est comparable en cela à celui des scientifiques expérimentaux.

C'est le théorème « central limite » qui dans son universalité, porte l'idée d'indépendance comme point fondamental de l'enseignement et de l'utilisation des statistiques, en justifiant l'emploi fréquent de la loi normale, centrée réduite.

Rôle des statistiques

L'enseignement des statistiques n'est pas fait pour démontrer ou infirmer l'existence du hasard, ou pour déceler sa présence. Les statistiques reposent sur le postulat initial que le monde est « probable » et partant de là, utilisent l'appareil théorique des probabilités.

Il est théoriquement possible qu'à partir de demain, il ne tombe plus une goutte de pluie sur Reims. Nous nous accordons pour penser que vraisemblablement il n'en sera rien. Nous vivons dans un monde où les événements de probabilité trop faible même subjective, ne se produisent pas, et nous agissons en conséquence. L'expérience ne nous à pas démenti ... jusqu'au jour où...

Le statisticien utilise des modèles probabilistes. Il ne peut jamais confirmer le modèle choisi. Il ne peut qu'infirmer ce modèle quand il constate une incompatibilité avec la série d'observations étudiées.

Vous avez dit "fréquentiste" ?

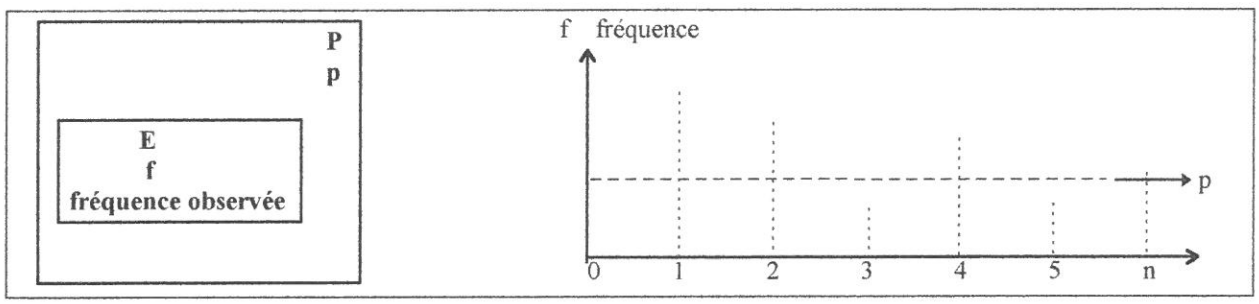
La rencontre entre "Pierre le géomètre" et "Jacques le fréquentiste"

Depuis de nombreuses années, Pierre le géomètre assurait un enseignement bien ciselé, construisant d'axiomes en théorèmes un édifice mathématique inébranlable. En particulier, en probabilités, il armait ses élèves et ses étudiants contre tout vice de raisonnement, les munissant de quelques bonnes lois et d'une trousse à outils de première intervention dans le cas fini renfermant dénombrement, arrangement et combinaison. Si le hasard prenait naturellement place dans l'énoncé de ses problèmes, il était chassé sans pitié dans leur résolution.

Mais, au détour d'un changement de programme, il rencontra Jacques le fréquentiste qui, d'emblée, fustigea sa vue démodée des probabilités en lui opposant la vision moderne : "l'approche fréquentiste". Bien que fortement déstabilisé par cette remise en cause, Pierre le géomètre fit l'effort d'écouter Jacques le fréquentiste. Il comprit assez vite que l'objet de cette nouvelle approche était le passage d'une fréquence observée, relevant du domaine de la statistique pour laquelle il avait un certain mépris, à sa transformation en probabilités, entachant par là ce beau modèle mathématique par cette intrusion externe et difficilement contrôlable !

En particulier, deux points le mettaient dans un fort embarras :

- la confusion, volontairement déclarée, entre deux types d'observation :
  * celle d'un échantillon **E** de la population **P** sur laquelle portait la probabilité étudiée,
  * celle de la répétition d'un événement aléatoire.



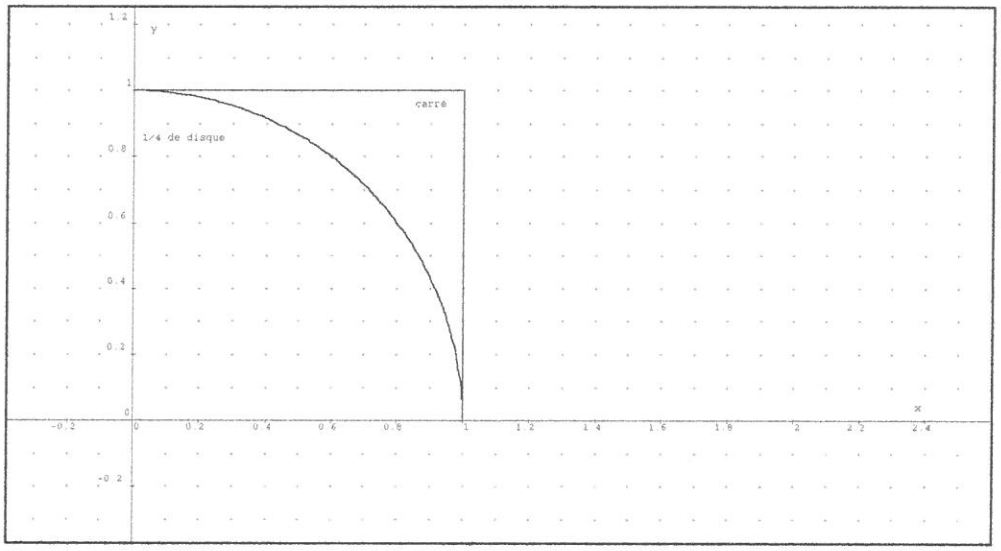
Si la première situation amène la question : "Ai-je choisi un bon échantillon ?" ou de façon plus mathématique "La fréquence du caractère étudié observée dans **E** est-elle proche de la fréquence de ce caractère dans la population totale **P** ?", la seconde pose la question "Ai-je bien attendu la stabilisation de la fréquence ?" ou de façon plus mathématique "La fréquence f_p choisie par arrêt au bout de p répétitions est-elle proche de la fréquence limite $\lim_{n\to+\infty} f_n$ que je choisis comme probabilité ?". Si les deux relèvent de la même problématique, à savoir la

difficulté ou l'impossibilité d'avoir accès à la population totale, soit qu'elle soit trop grande, soit a fortiori qu'elle soit infinie, elles se différencient fondamentalement au niveau du recueil de l'information, la première ne dépendant pas de l'ordre, la seconde résultant d'une chronologie.

- La présence systématique du hasard sur laquelle reposait ce nouvel édifice, présence ô combien déstabilisante pour un mathématicien habitué à la certitude et se trouvant dans l'impossibilité totale de mathématiser ce concept.

π par la méthode de Monte-Carlo

Plein de bonne volonté, Pierre le géomètre décida de mettre à l'épreuve de son enseignement les belles théories de Jacques le fréquentiste. Sur les conseils de ce dernier, il commença avec l'activité suivante :



"On tire un point "au hasard" dans le carré ci-dessus. Quelle est la probabilité qu'il appartienne au quart de disque ?".

Jacques le fréquentiste lui montra comment organiser l'activité de la classe autour de cette question :

* l'aide à la compréhension de l'énoncé,
* une première réflexion géométrique, le passage à la condition "M(x;y) appartient au quart de disque si et seulement si $0 \leq x \leq 1$, $0 \leq y \leq 1$ et $x^2 + y^2 \leq 1$",
* la découverte avec les élèves de la touche "RANDOM" de leur calculatrice (objet que jusque là Pierre le géomètre tenait en peu d'estime), leur fournissant "au hasard" des nombres compris entre 0 et 1,
* la consigne donnée à chaque élève de tester 20 couples de nombres "tirés au hasard" avec cette touche et de noter la fréquence de "tirs" dans le quart de disque,

* la collecte de toutes les informations de la classe en bénissant dans ce cas les classes à fort effectif,
* le produit par 4 de la fréquence obtenue donnant 3,1... (et dans les bons jours 3,14...),
* une tentative d'explication de ce miracle !

Si Pierre le géomètre dut reconnaître l'enthousiasme des élèves devant une telle activité, le peu d'écoute qu'il eut lorsqu'il voulut "expliquer" ce résultat le conduisit à émettre quelques réserves sur la portée scientifique d'une telle activité. Il en fit part à Jacques le fréquentiste :

"Si le but du problème est d'établir que la probabilité cherchée est $\frac{\pi}{4}$, alors ma bonne vieille géométrie et la loi de probabilité uniforme me conduisent immédiatement à ce résultat qui n'est autre que le quotient de l'aire du quart de disque par l'aire du carré".

"Si, au contraire, l'objectif est de montrer que l'on peut approcher ce résultat de manière fréquentiste, je vois alors quelques vices dans la démarche :

* Tout d'abord, ce passage du discret au continu est-il si évident que cela ?
* Plus grave, on ne part que du fini, la touche "RANDOM" se contentant de travailler avec trois chiffres après la virgule. On ne peut donc espérer trouver que le quotient du nombre de points à coordonnées de ce type intérieurs au disque par le nombre de points intérieurs au carré (et que dire des bords !).
* Enfin, si l'on considère que l'échantillon obtenu stabilise bien la fréquence cherchée, qu'a-t-on fait d'autre sinon vérifier que le générateur aléatoire de la calculatrice est bien aléatoire.

Jacques le fréquentiste balaya d'un "Tu as vu l'investissement des élèves ?" ces quelques remarques et l'encouragea à continuer dans la voie qu'il venait de s'ouvrir.

"Tu t'es vu quand t'as bu" ou "la parabole de l'ivrogne"

Pierre le géomètre décida de s'attaquer au problème du générateur aléatoire. Il lut beaucoup et tomba un jour sur un livre qu'il trouva merveilleux : "Les probabilités à l'école" de Maurice Glaymann et Tamas Varga. Et, au hasard d'une page, il découvrit une activité qui lui permit de donner du sens à sa question.

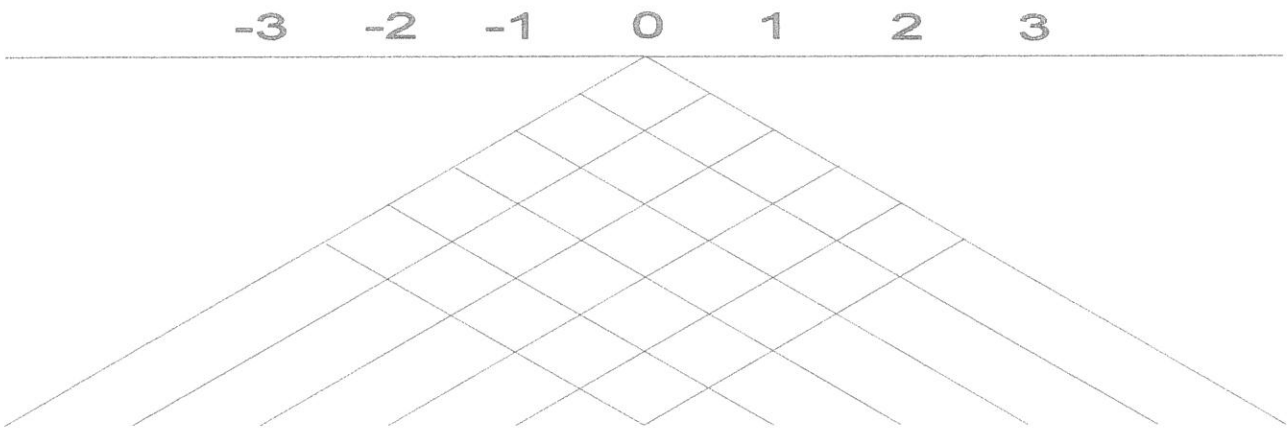
"Un ivrogne, pour rentrer chez lui, suit un chemin bordé d'arbres distants les uns des autres d'un mètre. Il se déplace d'un arbre à l'autre, mais à chaque déplacement, il oublie d'où il est venu et repart au hasard dans un sens ou dans l'autre."

Pierre le géomètre tenait enfin son générateur aléatoire, c'était l'ivrogne.

Il lut alors la suite de l'activité :

"Notre ivrogne tombe de sommeil au bout de n déplacements et s'endort au pied d'un arbre. Quelle est la distance moyenne entre l'arbre de départ et l'arbre où il s'endort ? (c'est-à-dire quelle est la moyenne des distances entre l'arbre de départ et les arbres où il a pu arriver au bout de n étapes ?)"

Il résolut assez rapidement le problème de la modélisation de la promenade de son ivrogne, s'émerveillant de retrouver le triangle de son idole mathématique : Pascal.

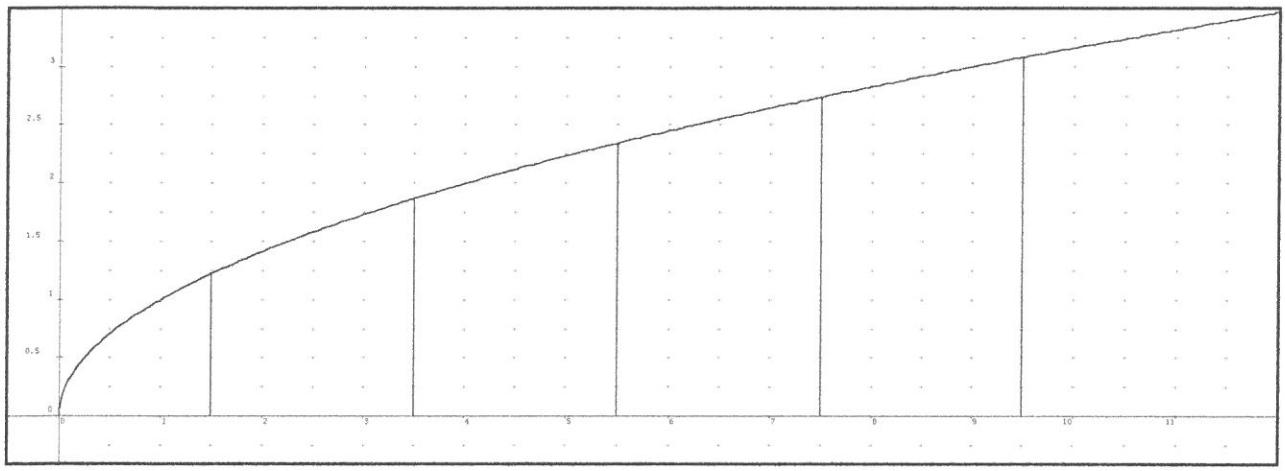


A coup de C_n^p, il établit alors le tableau suivant :

nombre d'étapes (n)	écart moyen (d)
1 ou 2	1
3 ou 4	1,5
5 ou 6	1,88
7 ou 8	2,19
9 ou 10	2,46
11 ou 12	2,71
13 ou 14	2,93
...	...

Il décida alors de représenter ce tableau par le graphique s'appuyant sur les points :

x	1,5	3,5	5,5	7,5	9,5	...
y	1	1,5	1,88	2,19	2,46	...



Force lui fut de constater que le chemin aléatoire de son ivrogne conduisait à un écart moyen tristement parabolique.

Poursuivant plus avant sa lecture, il découvrit que si son ivrogne avait une énergie infinie, la probabilité qu'il atteigne n'importe quel arbre pris au hasard était 1. Il apprit aussi que si cet ivrogne devait se déplacer non plus sur une droite mais dans un plan muni d'un quadrillage régulier d'arbres à un mètre les uns des autres, la probabilité d'atteindre n'importe quel arbre était toujours 1. Mais c'est avec stupeur qu'il dut admettre que ce résultat n'était plus vrai dans l'espace, la probabilité tombant alors à environ 0,35. Il remarqua avec plaisir que ce problème fut l'objet d'une épreuve du concours ESSEC 1987.

Triangle es-tu là ?

Fervent adepte des stages M.A.F.P.E.N., Pierre le géomètre évoluait beaucoup dans son enseignement. Il n'hésitait pas à mettre ses élèves en activité pour initier un nouveau concept mathématique. Il eut ainsi une idée de génie pour aborder l'inégalité triangulaire en $4^{ème}$: il revint avec un paquet de spaghettis, en distribua une poignée à chaque élève et leur proposa le travail suivant : vous prenez chacun des spaghettis, vous les coupez en trois morceaux au hasard, vous mesurez la longueur de chacun des morceaux et vous tentez de faire un triangle avec ces trois morceaux. L'état de la classe à la fin de la séquence, ainsi que l'inscription sournoisement inscrite dans son dos au tableau : "Oui mais des Panzanis !" le peinèrent beaucoup.

Il décida donc de revenir avec cette classe en troisième sur ce thème, mais en utilisant la calculatrice. "Tirez trois nombres au hasard avec votre calculatrice (touche "RANDOM"), multipliez chacun par 100 et essayez de faire un triangle avec les trois nombres obtenus, chacun d'eux représentant la longueur d'un côté. Il demanda à chaque élève de faire 20 essais. Comme il lui restait un quart d'heure, il proposa, sans idée préconçue de regrouper les informations pour déterminer la fréquence du triangle. Quelle ne fut pas sa surprise de trouver un résultat voisin de $\frac{1}{2}$!

Quelle ne fut pas sa gêne lorsque les élèves lui demandèrent si c'était vrai !

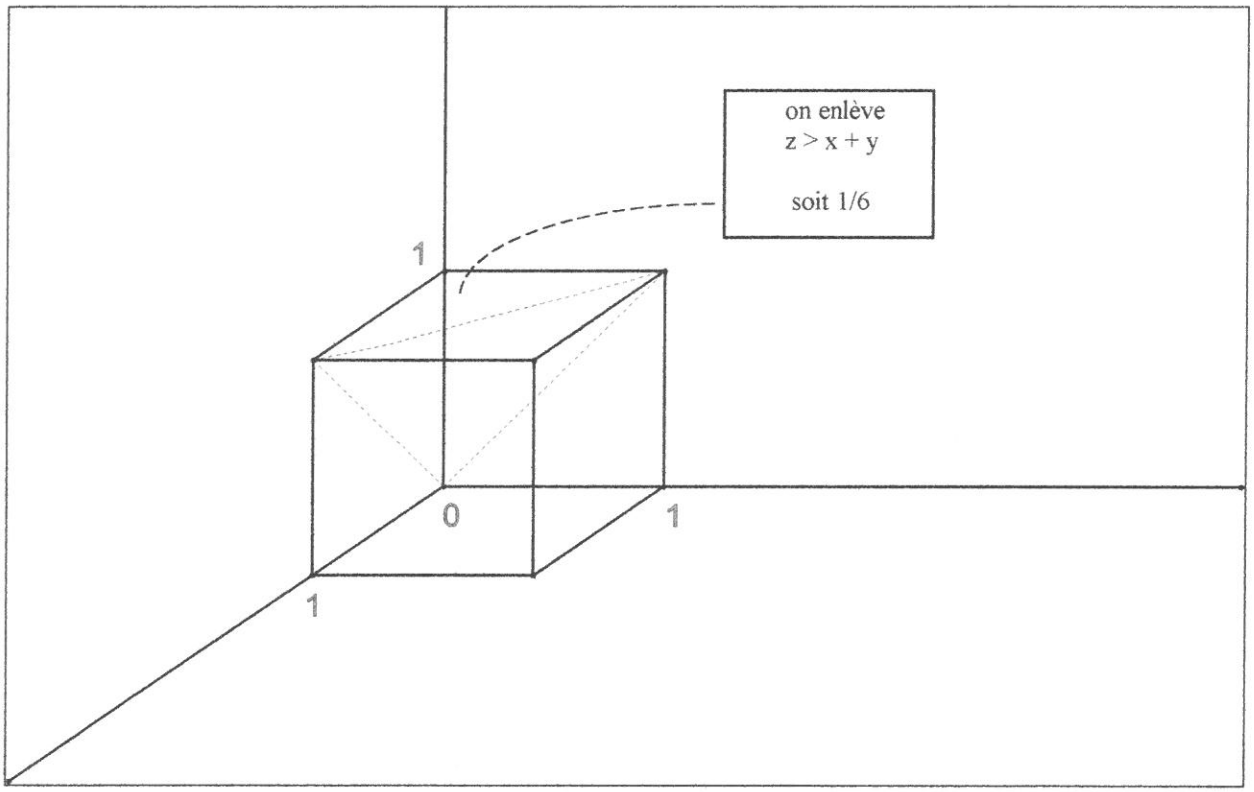
Il rencontra le plus vite possible Jacques le fréquentiste et lui fit part de son émoi. Celui-ci trouva tout de suite comment confirmer sa conjecture : "Teste avec d'autres classes, élargis ton échantillon !"

Et d'expériences en expériences, Pierre le géomètre confirmait cette fréquence de $\frac{1}{2}$, mais n'osait toujours pas la déclarer comme une probabilité.

Ne dormant presque plus, il remplissait des pages de triangles jusqu'au jour où le géomètre qui sommeillait en lui se réveilla :

"Tu as trois coordonnées x, y et z. Passe donc dans l'espace !"

Et comme pour l'approche de π par la méthode de Monte-Carlo, la solution lui apparut dans toute sa beauté géométrique :



"L'ensemble des triplets obtenus par ma calculatrice est représenté par le cube de côté 1. L'ensemble des triplets ne donnant pas un triangle s'obtient par 3 coupes de mon cube :

$$z > x + y \quad ; \quad x > y + z \quad ; \quad y > x + z.$$

Or chacun d'eux correspond au $\frac{1}{6}$ de mon cube avec une intersection vide de ces trois sous-espaces.

Donc ... !"

Tout fier de sa démonstration, Pierre le géomètre s'empressa de l'exposer à Jacques le fréquentiste. Celui-ci le prit assez mal et lui fit remarquer avec beaucoup de fiel : "Avec ta façon d'établir les choses, tu dois certainement pouvoir "calculer" la probabilité d'obtenir un triangle rectangle !".

Et lorsque Pierre le géomètre vit le résultat apparaître, Jacques le fréquentiste lui dit : "Comment peux-tu donc faire suer des générations d'élèves sur un objet géométrique dont la probabilité d'existence est nulle !".

Le jet de punaises

Poursuivant son avantage, et pour enfoncer définitivement le clou, Jacques le fréquentiste demanda à Pierre le géomètre : "Si tu jettes une punaise en l'air, lorsqu'elle retombe, qu'elle est la probabilité que ce soit sur la tête ?"

Devant la mine déconfite de Pierre le géomètre, il lui annonça que ce problème était actuellement l'objet d'une recherche fondamentale d'un I.R.E.M.

Pierre le géomètre n'hésita pas : il s'inscrivit à un stage proposé par cet I.R.E.M. sur ce thème. Il fut tout heureux d'être retenu. Il admira pendant la première matinée, l'approche expérimentale : chacun des stagiaires, à tour de rôle, lançait la même punaise et on comptabilisait la fréquence de "tombée sur la tête". Toutefois, un peu lassé, il eut l'outrecuidance, en fin de journée, d'émettre l'hypothèse, puisque l'on confondait échantillon et stabilisation de la fréquence, de lancer d'un coup 10 000 punaises et de compter ! Il s'attira alors les foudres de l'équipe d'animateurs-organisateurs-observateurs-comptabilisateurs qui lui firent remarquer assez sèchement que le stage durait trois jours !

Sus aux manuels

Si Pierre le géomètre dut bien reconnaître que pour ce fameux problème des punaises, seule l'approche fréquentiste permettait de décider d'une probabilité, il voulut toutefois se venger de la suffisance avec laquelle Jacques le fréquentiste le lui avait asséné. Il lui posa donc le problème :

"Je tire deux nombres entiers naturels non nuls au hasard. Quelle est la probabilité p qu'ils soient premiers entre eux ?"

Jacques le fréquentiste mit aussitôt en place une simulation reposant sur le tirage aléatoire successif de couples d'entiers, testant s'ils étaient premiers entre eux et proposa, au bout d'un certain temps un résultat.

Tout cela sous l'oeil amusé de Pierre le géomètre qui lui posa alors quelques questions :

"* Comment peux-tu être certain de l'existence d'une telle probabilité ?"

* Tu travailles avec un nombre fini d'entiers, qu'est-ce qui t'autorise à penser que la répartition des nombres premiers entre eux dans $[1;n] \times [1;n]$ est la même que dans $\mathbf{N}^* \times \mathbf{N}^*$?"

* Es-tu certain d'avoir assez attendu pour que la fréquence soit stabilisée ?"

Il reprit alors le problème avec un crayon et une feuille de papier et commença, d'un air docte :

"Je pourrais bien entendu te montrer l'existence de p, en procédant de la manière suivante :

Soit $\Omega_n = [1;n] \times [1;n]$, $Card(\Omega_n) = n^2$, $\Lambda_n = \{(x;y) \in \Omega_n \ / \ pgcd(x;y) = 1\}$, $Card(\Lambda_n) = q_n$.

Alors la probabilité que deux nombres x et y de [1;n] soient premiers entre eux est donc $p_n = \frac{q_n}{n^2}$.

On montre d'ailleurs que $p_n = \sum_{d=1}^{n} \frac{\mu(d)}{n^2} \left(E\left[\frac{n}{d}\right] \right)^2$ où μ est la fonction de Möbius et E désigne la fonction partie entière et tu peux alors vérifier, avec le "n" de tes tirages, si tu obtiens bien une fréquence voisine de p_n. Mais, comment peux-tu passer à $p = \lim_{n \to +\infty} p_n$."

Il continua d'un air bon enfant :

"Admettons l'existence de cette fameuse probabilité p. Je peux alors la calculer directement avec mes bonnes vieilles techniques de dénombrement :

soit $A_1 = \{(x;y) \in \mathbf{N}^* \times \mathbf{N}^* \ / \ pgcd(x;y) = 1\}$, alors p est la probabilité de l'événement A_1.

soit $A_2 = \{(x;y) \in \mathbf{N}^* \times \mathbf{N}^* \ / \ x$ et y sont divisibles par 2 et $pgcd(\frac{x}{2} ; \frac{y}{2}) = 1\}$, alors la probabilité de l'événement A_2 est $\frac{p}{4}$.

Puis de manière générale :

soit $A_n = \{(x;y) \in \mathbf{N}^* \times \mathbf{N}^* \ / \ x$ et y sont divisibles par n et $pgcd(\frac{x}{n} ; \frac{y}{n}) = 1\}$, alors la probabilité de l'événement A_n est $\frac{p}{n^2}$.

Les événements A_n sont incompatibles deux à deux et leur réunion est $\bigcup_{n=1}^{+\infty} A_n = \mathbf{N}^* \times \mathbf{N}^*$.

Donc $1 = P(\mathbf{N}^* \times \mathbf{N}^*) = P(\bigcup_{n=1}^{+\infty} A_n) = \sum_{n=1}^{+\infty} P(A_n) = \sum_{n=1}^{+\infty} \frac{p}{n^2} = p \sum_{n=1}^{+\infty} \frac{1}{n^2}$.

D'où $p = \frac{1}{\sum_{n=1}^{+\infty} \frac{1}{n^2}}$ ainsi $p = \frac{6}{\pi^2}$.

Il rit beaucoup de l'air déconfit de Jacques le fréquentiste et décida de pousser encore plus loin son avantage :

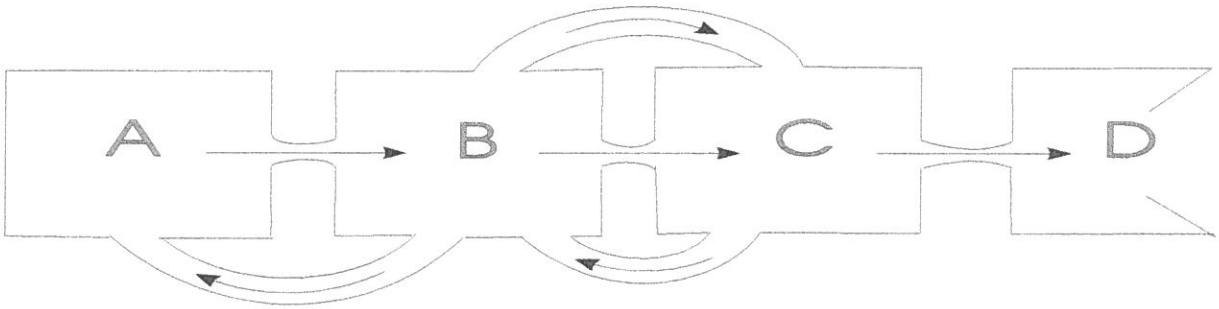
"π^2 est proche de 10, donc la probabilité cherchée est de l'ordre de 60 %. Je prends un manuel de $4^{ème}$ et choisis une page au hasard pleine de fractions. Utilisant les fractions pour couple d'entiers, je calcule la fréquence de celles solutions de mon problème et trouve 5 % !"

Jacques le fréquentiste trouva le procédé ignoble et fit remarquer, lisant la consigne en haut de la page : "Simplifiez les fractions suivantes." que tout ce qu'on pouvait conclure était que l'auteur dudit manuel s'était trompé dans 5 % de ses exercices !

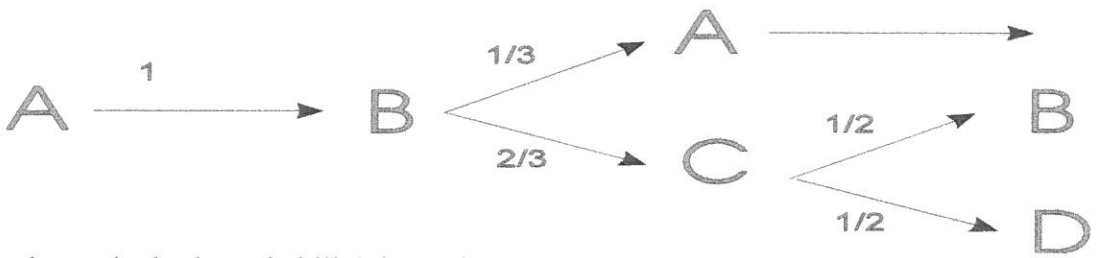
Ce à quoi Pierre le géomètre lui rétorqua que cet échantillon pris au hasard en valait bien un autre et qu'un non-matheux n'aurait eu a priori aucune raison de le rejeter comme étant aberrant !
Il s'en suivit une grande brouille entre les deux hommes.

Une souris aléatoire, un million de souris fréquentistes

Au bout de quelque temps, Pierre le géomètre eut un peu honte de l'attitude excessive qu'il avait eue avec Jacques le fréquentiste. Il le rencontra donc et lui dit :
"J'ai trouvé dans le merveilleux livre que je t'ai déjà signalé une activité qui va nous réconcilier."



Il lui proposa le problème du labyrinthe.
"Une souris aléatoire se trouve en A. Elle ne dispose que d'une issue qui la conduit en B. De B, elle dispose de trois issues, les deux premières la conduisant en C, la troisième la renvoyant en A. La souris choisit une de ces issues "au hasard" avec la même probabilité. Si elle arrive en C, elle dispose de deux issues qu'elle choisit encore au hasard avec la même probabilité, l'une la renvoyant en A, l'autre la conduisant à la sortie du labyrinthe D.
Au bout de quel nombre moyen d'étapes (c'est-à-dire passage d'une case à l'autre) la souris sort-elle du labyrinthe ?"
"Mon approche des probabilités me conduit à modéliser ce problème ainsi :



Je peux alors calculer la probabilité de sortir

- au bout de 3 étapes : $\frac{1}{3}$ $\left(1 \times \frac{2}{3} \times \frac{1}{2}\right)$
- au bout de 4 étapes : 0
- au bout de 5 étapes : $\frac{2}{9}$ $\left(\frac{1}{3} \times \frac{1}{3} + \frac{2}{3} \times \frac{1}{2} \times \frac{2}{3} \times \frac{1}{2}\right)$
- .....

- au bout de 2n étapes : 0

- au bout de (2n +1) étapes : $\frac{2^{n-1}}{3^n}$.

On vérifie bien que $\sum_{n=1}^{+\infty} \frac{2^{n-1}}{3^n} = \frac{1}{3}\sum_{n=0}^{+\infty}\left(\frac{2}{3}\right)^n = \frac{1}{3} \times 3 = 1.$

Je peux alors calculer l'espérance de sortie :

$$E = \sum_{n=1}^{+\infty} (2n+1)\frac{2^{n-1}}{3^n} = \sum_{n=1}^{+\infty} n\left(\frac{2}{3}\right)^n + \frac{1}{2}\sum_{n=1}^{+\infty}\left(\frac{2}{3}\right)^n = \frac{2}{3} \times \sum_{n=1}^{+\infty} n\left(\frac{2}{3}\right)^{n-1} + \frac{1}{2}\sum_{n=1}^{+\infty}\left(\frac{2}{3}\right)^n = \frac{2}{3} \times 9 + 1 = 7.$$

(Pour montrer que $\sum_{n=1}^{+\infty} n\left(\frac{2}{3}\right)^{n-1} = 9$, utiliser, par exemple, la dérivée de la série entière $\sum_{n=1}^{+\infty} x^n$.)

J'en conclus que le nombre moyen d'étapes pour sortir est N = 7."

Jacques le fréquentiste fut très intéressé par le problème et par la façon très élégante dont Pierre le géomètre en était venu à bout. Mais il se retrouva davantage dans la deuxième façon de l'aborder.

"Faisons entrer un million de souris fréquentistes dans le labyrinthe. Elles se répartissent de façon aléatoire dans les couloirs, on peut alors compter à la fin de chaque étape combien sont dans chacune des cases. Pour éviter tout problème avec la S.P.A., il est interdit de couper les souris en 3 ou en 2. On arrondira donc les effectifs pour éviter la troncature des souris, ce qui permettra en outre d'achever le problème en un nombre fini d'étapes.

La surveillance des différentes cases conduit au tableau suivant :

Etape	A	B	C	D
1	0	1 000 000	0	0
2	333 333	0	666 667	0
3	0	666 667	0	333 333
4	222 222	0	444 444	0
5	0	444 444	0	222 222
...	...	...	...	...
20	8 671	0	17 342	0
21	0	17 342	0	8 671
...	...	...	...	...
39	451	451	0	226
40	0	0	301	0
...	...	...	...	...

Et au bout de 80 étapes, toutes les souris sont sorties. On calcule alors la moyenne de sortie :

$N = \frac{333\,333 \times 3 + 222\,222 \times 5 + \ldots + 8671 \times 21 + \ldots + 226 \times 39 + \ldots}{80}$; et on trouve $N \approx 7$."

Ce problème réconcilia les deux hommes qui rirent de leur brouille passée. Ils devinrent des amis inséparables.

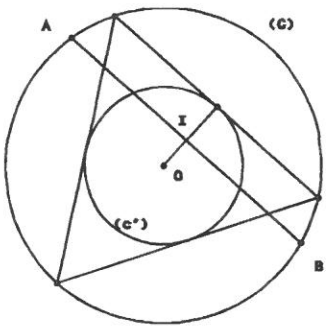
La rencontre avec Bertrand le paradoxal

On les vit beaucoup ensemble dans de nombreuses réunions, colloques, universités d'été ... C'est lors d'un séminaire qu'ils firent la rencontre de Bertrand le paradoxal. Ils lui racontèrent l'histoire de leur difficile amitié et la sérénité qu'ils avaient acquise en statistiques et en probabilité. Bertrand le paradoxal leur posa alors le problème suivant :

"On considère un triangle équilatéral de côté a. Soit C(O,R) le cercle circonscrit à ce triangle. On tire "au hasard" une corde [AB] de C. Quelle est la probabilité p que AB > a ?"

Et il enchaîna : "Je vous propose trois solutions."

Première méthode :



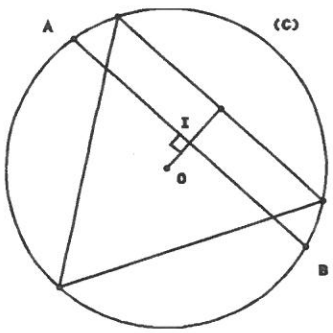
Je considère le cercle $C'(O,\frac{R}{2})$. Soit I le milieu de [AB].

Le problème devient :

"Quelle est la probabilité que I soit intérieur à C' ?"

Alors $p = \frac{1}{4}$ (rapport des aires).

Deuxième méthode :



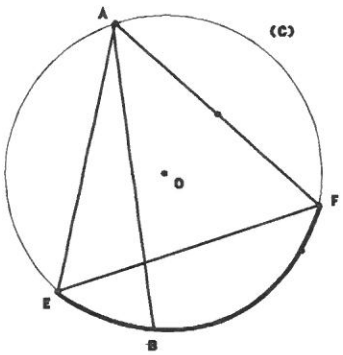
Je calcule la distance d de (AB) à O.

Le problème devient :

"Quelle est la probabilité que cette distance d soit inférieure à la distance des côtés du triangle à O, soit $\frac{R}{2}$?

Alors $p = \frac{1}{2}$ (rapport des distances).

Troisième méthode :



A ayant été tiré au hasard sur C, je construis le triangle AEF équilatéral inscrit dans C.

Le problème devient :

"Quelle est la probabilité que B soit sur l'arc intercepté par $E\hat{A}F$?"

Alors $p = \frac{1}{3}$ (rapport des arcs).

Et devant leurs mines abasourdies, Bertrand le paradoxal ajouta :

"Le choix du modèle, messieurs, le choix du modèle !"

Pierre le géomètre et Jacques le fréquentiste rentrèrent tristes de ce séminaire, trouvant que vraiment les probabilités étaient une science bien déstabilisante. D'un commun accord, ils décidèrent de renvoyer son enseignement en fin d'année scolaire.

STATISTIQUES ET PROBABILITES

-

PROBABILITES ET STATISTIQUES

Des lois de probabilité

A. Préliminaire

I. Probabilité

Soit Ω un univers non vide et $P(\Omega)$ l'ensemble des événements associé à Ω.

Une probabilité sur $(\Omega, P(\Omega))$ [ou "sur Ω"] est une application :

$$\begin{aligned} P : P(\Omega) &\longrightarrow [0,1] \\ A &\longmapsto P(A) \end{aligned}$$

satisfaisant les axiomes de Kolmogorov (1937) :

1. $P(\Omega) = 1$

2. Si $A \cap B = \varnothing$ alors $P(A \cup B) = P(A) + P(B)$ (P est une application additive)

Et dans le cas où Ω est fini :

3. Si $(A_i)_{i \in N}$ est une famille dénombrable d'événements deux à deux incompatibles,

alors $P\left(\bigcup_{i \in N} A_i\right) = \sum_{i \in N} P(A_i)$ (P est une application σ-additive)

II. Variable aléatoire

Soit P une probabilité définie sur $(\Omega, P(\Omega))$

Une variable aléatoire définie sur Ω est une application $X : \Omega \longmapsto \mathbb{R}$.

III. Probabilité image de P par X

C'est l'application : $\begin{aligned} P' : P(X(\Omega)) &\longrightarrow [0,1] \\ B &\longmapsto P'(B) = P(X^{-1}(B)) \end{aligned}$

On note simplement par la suite : $P(X \in B) = P(X^{-1}(B))$.

IV. Fonction de répartition

Soit X une variable aléatoire définie sur Ω. La fonction de répartition de X est l'application

$$\begin{aligned} F : \mathbb{R} &\longrightarrow [0,1] \\ x &\longmapsto f(x) = P(X \leq x) = P(X^{-1}(]-\infty, x])) \end{aligned}$$

Attention à la définition anglo-saxonne, $F(x) = P(X \geq x)$, surtout pour utiliser des répartitions de la librairie des calculatrices Casio FX-850 P (binomiale, Poisson, normale, ..., 6210 à 6330 LIB).

Propriétés :

1. F est croissante sur $\mathbb{R}$

2. F est continue à droite en tout point de $\mathbb{R}$

3. $\lim_{-\infty} F = 0$ et $\lim_{+\infty} F = 1$

Remarques :

F est monotone et continue à droite, elle présente donc un ensemble au plus dénombrable de points de discontinuité.

Réciproquement, toute fonction vérifiant les conditions **1.**, **2.** et **3.** est la fonction de répartition d'une variable aléatoire.

V. Les trois types de variables aléatoires

1. $X(\Omega)$ est un ensemble fini : X est une ***variable aléatoire discrète finie***.

Exemple :

On lance deux dés. Soit $E = \{1, 2, 3, 4, 5, 6\}$. On a $\Omega = E^2 = \{(1,1), (1,2), \ldots, (6,6)\}$.

On se place en situation d'équiprobabilité, on définit donc une probabilité sur $(\Omega, P(\Omega))$ par :

$$P : P(\Omega) \longrightarrow [0,1]$$
$$A \longmapsto \frac{\text{Card}A}{\text{Card}\,\Omega}$$

La probabilité de l'événement A est le rapport du nombre de cas favorables à la réalisation de A au nombre de tous les cas possibles.

Soit X la variable aléatoire mesurant la somme des points marqués par les deux dés.

On a alors $X(\Omega) = \{2, 3, \ldots, 12\} = [\![2, 12]\!]$

$$P' : [\![2, 12]\!] \longrightarrow [0,1]$$
$$B \longmapsto P'(B) = P(X^{-1}(B)) = P(\{\omega \in \Omega \mid X(\omega) \in B\})$$

Ainsi, par exemple :

$P'(\{4\}) = P(\{(1,3);(3,1);(2,2)\}) = \frac{3}{36}$; soit avec la notation simplifiée $P(X = 4) = \frac{3}{36}$;

$F(4) = P(X \leq 4) = P([X = 2] \cup [X = 3] \cup [X = 4]) = P(X = 2) + P(X = 3) + P(X = 4)$

$= \frac{1}{36} + \frac{2}{36} + \frac{3}{36} = \frac{1}{6}$.

2. $X(\Omega)$ est un sous-ensemble dénombrable de $\mathbb{R}$

X est une ***variable aléatoire discrète dénombrable*** (ou discrète infinie).

Exemple :

On lance un dé. Soit X la variable aléatoire mesurant le nombre de jets jusqu'à la réalisation de l'événement : "le dé donne un as". $X(\Omega) = \mathbb{N}^*$.

Soit A_i l'événement "le $i^{ème}$ lancer donne un as".

On obtient alors : quel que soit $n \in \mathbb{N}^*$, $P(X = n) = P(\overline{A_1} \cap \overline{A_2} \cap \ldots \cap \overline{A_{n-1}} \cap A_n)$ et si l'on formule l'hypothèse que le dé est sans mémoire, c'est-à-dire que chaque résultat n'a pas d'influence sur les précédents, ni sur les suivants, alors les événements A_i sont mutuellement indépendants. On en déduit : $P(X = n) = P(\overline{A_1}) \times P(\overline{A_2}) \times \ldots \times P(\overline{A_{n-1}}) \times P(A_n) = \left(\frac{5}{6}\right)^{n-1} \times \frac{1}{6}$

3. $X(\Omega)$ est une réunion d'intervalles de $\mathbb{R}$ non réduite à un point et la fonction de répartition de X est définie sur $\mathbb{R}$ par $F(x) = \int_{-\infty}^{x} f(t)\,dt$ où f est une fonction positive, admettant un nombre fini de points de discontinuité, et telle que $\int_{-\infty}^{+\infty} f(t)\,dt = 1$.

X est alors une ***variable aléatoire à densité*** ou encore ***variable aléatoire continue*** (ou plus exactement ***absolument continue)***. La fonction f est ***une*** densité de probabilité de X.

Conséquences pratiques :

1. $\forall x \in \mathbb{R}$, $P(X < x) = P(X \leq x) = F(x) = \int_{-\infty}^{x} f(t)\,dt$.

2. $\forall (a, b) \in \mathbb{R}^2$, $P(a \leq X \leq b) = F(b) - F(a) = \int_{a}^{b} f(t)\,dt$.

3. $\forall x \in \mathbb{R}$, $P(X = x) = 0$

Interprétation graphique :

Soit X une variable aléatoire de densité de probabilité f et C_f la courbe représentative de f tracée dans un repère orthogonal. L'aire totale "sous la courbe" C_f vaut une unité d'aire.

$\forall (a, b) \in \mathbb{R}^2$, $P(a \leq X \leq b)$ est l'aire du domaine de frontières l'axe des abscisses, la courbe C_f, les droites d'équations $x = a$ et $x = b$.

$\forall a \in \mathbb{R}$, $P(X = a)$ est l'aire du segment porté par la droite d'équation $x = a$, dont les extrémités sont sur l'axe des abscisses et la courbe C_f : c'est un ensemble de mesure nulle.

Exemple 1 :

Soit la fonction F définie par : $\begin{cases} F(x) = 0 & \text{si } x < 0 \\ F(x) = 0{,}002 \int_0^x e^{-0{,}002t}\,dt & \text{si } x \geq 0 \end{cases}$.

On se propose de vérifier que F est la répartition d'une variable aléatoire X. La durée de vie d'un composant électronique, exprimée en jours, est mesurée par cette variable aléatoire X. Calculer la probabilité que la durée de vie d'un tel composant pris au hasard dans la production n'excède pas 400 jours. Calculer la durée de vie moyenne d'un tel composant électronique, c'est-à-dire l'espérance mathématique de la variable aléatoire X.

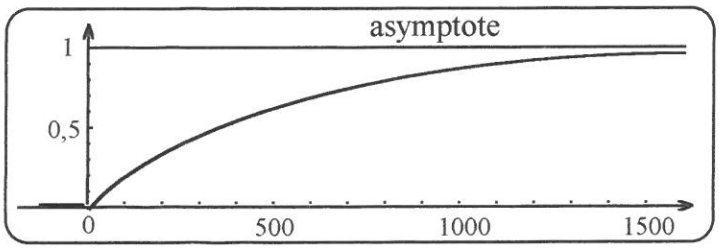
◆ *Solution* :

F est dérivable, sa dérivée est définie par : $\begin{cases} F'(x) = 0 & \text{si } x < 0 \\ F'(x) = 0{,}002\, e^{-0{,}002x} & \text{si } x \geq 0 \end{cases}$.

$F' \geq 0$ sur $\mathbb{R}$ donc F est croissante ; $F(x) = 0$ si $x < 0$ implique $\lim\limits_{-\infty} F = 0$.

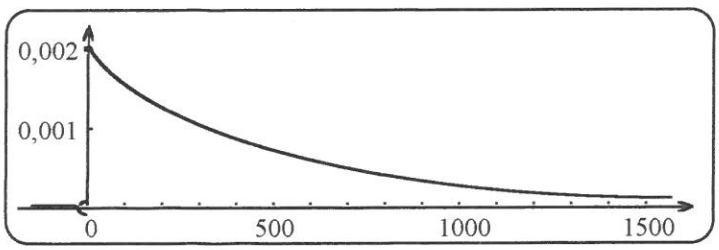
Si $x \geq 0$, $F(x) = [-e^{-0{,}002t}]_0^x = 1 - e^{-0{,}002x}$, donc $\lim\limits_{+\infty} F = 1$.

Ce qui établit que F est la fonction de répartition d'une variable aléatoire X.



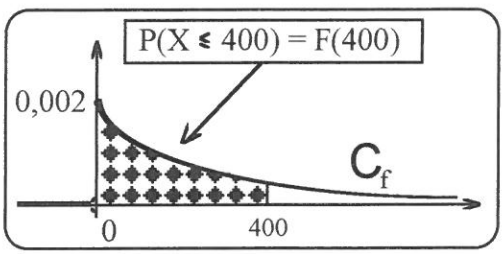
Représentation graphique de la fonction de répartition F de X.

La fonction dérivée F' de la fonction de répartition F est la fonction densité de probabilité f de X.



Représentation graphique de la fonction densité de probabilité f de X.

$P(X \leq 400) = F(400) = [-e^{-0{,}002t}]_0^{400} = 1 - e^{-0{,}8} \approx 0{,}55.$



$E(X) = \int_{-\infty}^{+\infty} x f(x)\, dx = 0{,}002 \int_0^{+\infty} x\, e^{-0{,}002x}\, dx$

Une intégration par parties donne alors :

$$E(X) = [-x\, e^{-0{,}002x}]_0^{+\infty} + \int_0^{+\infty} e^{-0{,}002x}\, dx = \left[\frac{-e^{-0{,}002x}}{0{,}002}\right]_0^{+\infty} = \frac{1}{0{,}002} = 500.$$

❖ ❖ ❖

***Exemple 2* :**

Soit F la fonction définie par : $\begin{cases} F(x) = 0 & \text{si } x < 0 \\ F(x) = 1 - \frac{2}{3} e^{-\frac{2}{3}x} - \frac{1}{3} e^{-\frac{1}{3}x} & \text{si } x \geq 0 \end{cases}$

◆ *Solution* :

Il est facile de vérifier que F est continue et croissante, que $\lim_{-\infty} F = 0$ et $\lim_{+\infty} F = 1$.

La densité de probabilité f de la variable aléatoire X est définie pour tout x de $\mathbb{R}$ par $f(x) = F'(x)$.

On en déduit que : $f(x) = 0$ si $x < 0$ et que : $f(x) = \frac{4}{9} e^{-\frac{2}{3}x} + \frac{1}{9} e^{-\frac{1}{3}x}$ si $x \geq 0$.

Calcul de l'espérance mathématique de X : (intégration par parties, puis calcul de limites usuelles)

$$E(X) = \int_{-\infty}^{+\infty} t f(t)\, dt = \int_{0}^{+\infty} t\, f(t)\, dt = \lim_{x \to +\infty} \int_{0}^{x} t\, f(t)\, dt = 2 .$$

$$P_{[X > 1]}(X \leq 2) = \frac{P(1 < X \leq 2)}{P(X > 1)} = \frac{F(2) - F(1)}{1 - F(1)} \approx 0{,}4031.$$

❖ ❖ ❖

***Exemple 3* :**

Soit f la fonction définie par : $\begin{cases} f(x) = 0 & \text{si } x < 0 \text{ ou } x > 1 \\ f(x) = \frac{4}{3}\sqrt[3]{1 - x} & \text{si } 0 \leq x \leq 1 \end{cases}$

Montrer que f est la densité de probabilité d'une variable aléatoire X. Déterminer la fonction de répartition F de cette variable aléatoire X. Calculer l'espérance de la variable aléatoire X. Calculer la probabilité de l'événement $[0{,}2 < Y \leq 1{,}2]$.

◆ *Solution* :

La fonction f est positive et continue en tout point de $\mathbb{R}^*$.

$$\int_{-\infty}^{+\infty} f(x)\, dx = \int_{0}^{1} f(x)\, dx = \left[-(1 - x)^{\frac{4}{3}} \right]_{0}^{1} = 1.$$

La fonction f est donc la densité de probabilité d'une variable aléatoire X.

La fonction de répartition de X est une primitive de f. On en déduit que :

- F est constante sur $]-\infty;0[$ et comme sa limite en $-\infty$ est nulle, on a : $F(x) = 0$ si $x < 0$;
- F est constante sur $]1;+\infty[$ et comme sa limite en $+\infty$ est 1, on a : $F(x) = 1$ si $x > 1$;
- $\exists\, \lambda \in \mathbb{R}, \forall\, x \in [0,1],\ F(x) = -(1 - x)^{\frac{4}{3}} + \lambda$, la continuité de F impose $\lambda = 1$.

Calcul de l'espérance de la variable aléatoire X :

$$\int_{-\infty}^{+\infty} x\, f(x)\, dx = \int_{0}^{1} x\, f(x)\, dx = \frac{3}{7} \qquad \text{(obtenu en intégrant par parties).}$$

Calcul de la probabilité de l'événement $[0{,}2 < Y \leq 1{,}2]$:

$$P(0{,}2 < Y \leq 1{,}2) = F(1{,}2) - F(0{,}2) = 1 - \left(1 - 0{,}8^{\frac{4}{3}}\right) = 0{,}8^{\frac{4}{3}} \approx 0{,}743 .$$

❖ ❖ ❖

Exemple 4 : *(d'après B.T.S. Maintenance et Exploitation des Matériels Aéronautiques 1993).*

La durée de vie d'un tube de radio, exprimée en heures, est mesurée par une variable aléatoire T, dont la densité de probabilité est définie par :

$f(t) = 0$ si $t < 0$; $f(t) = \frac{1}{a} e^{-\frac{t}{a}}$ si $t \geq 0$, a étant un paramètre réel strictement positif.

Etudier les variations de la fonction f, donner son tableau de variation et tracer sa courbe représentative dans un repère orthogonal.

Soit t_0 un réel de $[0,+\infty[$. Calculer $\int_0^{t_0} f(t)\,dt$, puis $\lim\limits_{t_0 \to +\infty} \int_0^{t_0} f(t)\,dt$.

La probabilité que le tube soit hors service avant un nombre d'heures égal à t_0 est :

$$P(T < t_0) = \int_0^{t_0} f(t)\,dt$$

Calculer la probabilité de survie après t_0 heures, c'est-à-dire $P(T \geq t_0)$.

A l'aide d'une intégration par parties, calculer en fonction de a et de l'entier naturel n, $\int_0^n t\,f(t)\,dt$.

En déduire, en fonction de a, la durée de vie moyenne d'un tube définie par $E(T) = \lim\limits_{n \to +\infty} \int_0^{t_0} t\,f(t)\,dt$

Déterminer a sachant que la durée de vie moyenne d'un tube est de 1 000 heures.

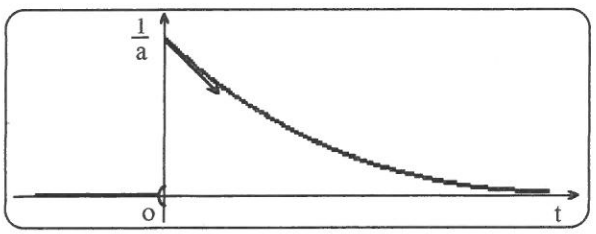
Au bout de combien de temps faut-il prévoir de remplacer un tube si l'on considère qu'il ne faut pas dépasser une probabilité de survie de 0,3 ?

Quatre tubes de ce type sont utilisés en parallèle dans un appareil électrique. Calculer la probabilité que cet appareil soit hors service en moins de 2 000 heures.

◆ *Solution* : •

Etude des variations de la fonction f : si $t < 0$ alors $f'(t) = 0$ et si $t \geq 0$ alors $f'(t) = -\frac{1}{a^2} e^{-\frac{t}{a}}$.

t	$-\infty$		0		$+\infty$
f'(t)	0		‖	-	
f(t)	0	⟶	0	$\frac{1}{a}$ ↘	0



• $\int_0^{t_0} f(t)\,dt = \left[-e^{-\frac{t}{a}} \right]_0^{t_0} = -e^{-\frac{t}{a}} + 1$ d'où $\lim\limits_{t_0 \to +\infty} \int_0^{t_0} f(t)\,dt = 1$.

• Probabilité de survie d'un tube après t_0 heures : $P(T \geq t_0) = 1 - P(T < t_0) = 1 - \int_0^{t_0} f(t)\,dt = e^{-\frac{t}{a}}$.

- Calcul de l'espérance mathématique de T : on pose $\begin{cases} u(t) = t \\ v'(t) = f(t) \end{cases}$ et $\begin{cases} u'(t) = 1 \\ v(t) = -e^{-\frac{t}{a}} \end{cases}$

d'où : $\int_0^n t\, f(t)\, dt = \left[-t\, e^{-\frac{t}{a}}\right]_0^n + \int_0^n e^{-\frac{t}{a}}\, dt = -n\, e^{-\frac{n}{a}} - a\left[e^{-\frac{t}{a}}\right]_0^n = -n\, e^{-\frac{n}{a}} - a\left(e^{-\frac{n}{a}} - 1\right)$

$$\int_0^n t\, f(t)\, dt = -n\, e^{-\frac{n}{a}} - a\, e^{-\frac{n}{a}} + a\,.$$

Soit $u = -\frac{n}{a}$, on a alors : $\lim\limits_{n\to+\infty}\left(-n\, e^{-\frac{n}{a}}\right) = \lim\limits_{u\to-\infty}(a\, u\, e^u) = 0$. Or $\lim\limits_{n\to+\infty}\left(e^{-\frac{n}{a}}\right) = 0$.

La durée de vie moyenne d'un tube est $E(T) = \lim\limits_{n\to+\infty} \int_0^{t_0} t\, f(t)\, dt = a$.

On en déduit que $a = 1\,000$.

- Calcul du temps t_0 au bout duquel la probabilité de survie d'un tube est inférieure à 0,3.

Un calcul précédent montre que $P(T \geq t_0) < 0{,}3$ équivaut à $e^{-\frac{t_0}{a}} < 0{,}3$.

On en déduit, avec $a = 1\,000$: $-\frac{t_0}{1\,000} < \ln 0{,}3$ soit encore $t_0 > -1\,000 \ln 0{,}3$.

La calculatrice donne $-1\,000 \ln 0{,}3 = 1\,903{,}97\ldots$

Il faut donc remplacer les tubes après 1 903 heures de fonctionnement.

- Soit A_k , pour $k \in \{1, 2, 3, 4\}$, l'événement "le $k^{ième}$ tube est hors service en moins de 2 000 heures". On suppose que ces événements sont indépendants et puisque les tubes sont montés en parallèle, l'événement "l'appareil est hors service en moins de 2 000 heures" est $A_1 \cap A_2 \cap A_3 \cap A_4$.

La probabilité de cet événement est donc :

$$P(A_1 \cap A_2 \cap A_3 \cap A_4) = P(A_1) \times P(A_2) \times P(A_3) \times P(A_4) = \left(P(T < 2\,000)\right)^4$$
$$= \left[\int_0^{2\,000} f(t)\, dt\right]^4 = \left(1 - e^{-2}\right)^4$$

La probabilité que l'appareil soit hors service en moins de 2 000 heures est environ 0,56.

❖ ❖ ❖

B. Lois de probabilité usuelles

I. Lois discrètes

1°. Introduction

La **loi de probabilité** (ou *distribution*) d'une variable aléatoire discrète X est l'application :

$$\begin{array}{ccc} X(\Omega) & \longrightarrow & [0,1] \\ k & \longmapsto & P(X = k) \end{array}$$

L'**espérance mathématique** de X est, lorsqu'elle existe, le nombre réel :

$$E(X) = \sum_{x_i \in X(\Omega)} x_i P(X = x_i)$$

La **variance** de X, lorsqu'elle existe, est le nombre réel $\boxed{V(X) = E([X - E(X)]^2)}$

Théorème de König-Huygens : $\boxed{V(X) = E(X^2) - [E(X)]^2}$

On note $\Sigma = \sum_{x_i \in X(\Omega)}$ et $p_i = P(X = x_i)$. On a alors $E(X) = \Sigma\, x_i\, p_i$ et $E(X^2) = \Sigma\, x_i^2\, p_i$ d'où :

$$V(X) = \Sigma\, [x_i - E(X)]^2\, p_i = \Sigma\, (x_i^2 - 2\, x_i\, E(X) + [E(X)]^2)\, p_i$$

$$V(X) = \Sigma\, (x_i^2\, p_i) - 2\, \Sigma\, x_i\, E(X)\, p_i + \Sigma\, [E(X)]^2\, p_i = E(X^2) - 2\, E(X)\, \Sigma\, x_i\, p_i + [E(X)]^2\, \Sigma\, p_i$$

$$V(X) = E(X^2) - 2\, [E(X)]^2 + [E(X)]^2 = E(X^2) - [E(X)]^2.$$

L'**écart type** de X est la racine carrée de la variance : $\sigma(X) = \sqrt{V(X)}$.

Précision orthographique : "écart type" s'écrit sans trait d'union.

Exemples :

❍ *Résultats préliminaires :*

Pour tout $x \in]0, 1[$, on a $\sum_{k=1}^{n} x^k = x\,\dfrac{1 - x^n}{1 - x}$, et par suite : $\lim\limits_{n \to +\infty} \sum_{k=1}^{n} x^k = \dfrac{x}{1 - x}$.

Par dérivation, on obtient une série convergente et $\lim\limits_{n \to +\infty} \sum_{k=1}^{n} k\, x^{k-1} = \dfrac{1}{(1 - x)^2}$.

La dérivée seconde donne $\lim\limits_{n \to +\infty} \sum_{k=1}^{n} k\,(k - 1)\, x^{k-2} = \dfrac{2}{(1 - x)^3}$, qui permet d'obtenir :

$$\lim_{n \to +\infty} \sum_{k=1}^{n} k^2\, x^k = \frac{x\,(x + 1)}{(1 - x)^3}.$$

***Exemple 1* :**

Soit X la variable aléatoire mesurant le nombre de lancers d'un dé jusqu'à la réalisation de l'événement : "le dé donne un as".

On a obtenu précédemment : $\forall\, n \in \mathbb{N}^*,\ P(X = n) = \left(\dfrac{5}{6}\right)^{n-1} \times \dfrac{1}{6}$.

L'espérance mathématique est alors $E(X) = \sum_{n\in N^*} n\,P(X=n) = \lim_{n\to+\infty} \sum_{k=1}^{n} k\left(\frac{5}{6}\right)^{k-1} \times \frac{1}{6} = \frac{1}{\left(1-\frac{5}{6}\right)^2} \times \frac{1}{6}$ soit

$E(X) = 6$. On peut espérer mathématiquement que l'on obtiendra un as en moyenne en six coups.

La variance de X est $V(X) = \sum_{n\in N^*} n^2\,P(X=n) - 6^2 = \lim_{n\to+\infty} \sum_{k=1}^{n} k^2\left(\frac{5}{6}\right)^{k-1} \times \frac{1}{6} - 6^2$

D'où $V(X) = \lim_{n\to+\infty} \frac{1}{5}\sum_{k=1}^{n} k^2\left(\frac{5}{6}\right)^{k} - 6^2 = \frac{1}{5} \times \frac{\frac{5}{6}\left(1+\frac{5}{6}\right)}{\left(1-\frac{5}{6}\right)^3} - 6^2 = 11 \times 6 - 6^2 = 30$.

❖ ❖ ❖

Exemple 2 :

Soit Y la variable aléatoire mesurant le gain d'un joueur auquel on attribue 6^n francs lorsque le dé donne un as pour la première fois au $n^{ième}$ lancer. On a $P(Y = 6^n) = P(X = n)$, où X est la variable aléatoire de l'exemple précédent. $\sum_{n\in N^*} 6^n\,P(Y=6^n) = \lim_{n\to+\infty} \sum_{k=1}^{n} 6^k\left(\frac{5}{6}\right)^{k-1} \times \frac{1}{6} = \lim_{n\to+\infty} \sum_{k=1}^{n} 5^{k-1} = +\infty$.
La variable aléatoire Y n'a pas d'espérance mathématique.

Propriétés :

Quels que soient les nombres réels a et b et les variables aléatoires discrètes X et Y, on a :

$\boxed{E(a\,X + b) = a\,E(X) + b}$ $\qquad$ $\boxed{E(X + Y) = E(X) + E(Y)}$

$\boxed{V(a\,X + b) = a^2\,V(X)}$ $\qquad$ $\boxed{\sigma(a\,X + b) = |a|\,s(X)}$

Les variables aléatoires X et Y sont *indépendantes* si et seulement si

$$\forall\, x_i \in X(\Omega) \text{ et } \forall\, y_j \in Y(\Omega),\ P([X = x_i] \cap [Y = y_j]) = P(X = x_i) \times P(Y = y_j).$$

Si les variables aléatoires X et Y sont indépendantes, on a :

$\boxed{E(X\,Y) = E(X)\,E(Y)}$ $\qquad$ $\boxed{V(X + Y) = V(X) + V(Y)}$ $\qquad$ $\boxed{\sigma(X + Y) = \sqrt{\sigma(X)^2 + \sigma(Y)^2}}$

2°. Loi uniforme discrète

La variable aléatoire X suit la loi uniforme U_n, ce que l'on note $X \hookrightarrow U_n$, lorsque :

$$X(\Omega) = \{x_1, x_2, \ldots, x_n\} \text{ et quel que soit } x_k \in X(\Omega),\ P(x_k) = \frac{1}{n}.$$

Exemple :

On lance un dé, la variable aléatoire X donne le résultat : X suit la loi U_6.

$X(\Omega) = \{1, 2, 3, 4, 5, 6\}$ et quel que soit $n \in X(\Omega)$, $P(n) = \frac{1}{6}$.

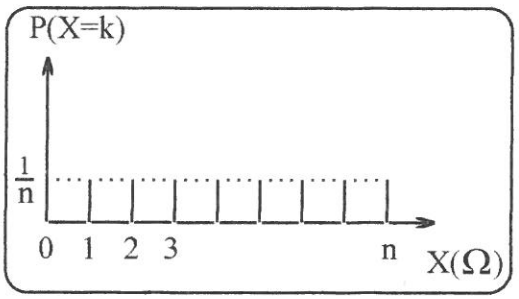
$$E(X) = \sum_{k=1}^{6} k\,P(X=k) = \sum_{k=1}^{6} \frac{k}{6} = \frac{1}{6}\sum_{k=1}^{6} k = \frac{1}{6}\frac{6\times 7}{2} = 3{,}5.$$

$$V(X) = \sum_{k=1}^{6} k^2\,P(X=k) - 3{,}5^2 = \sum_{k=1}^{6} \frac{k^2}{6} - 3{,}5^2 = \frac{1}{6}\sum_{k=1}^{6} k^2 - 3{,}5^2 = \frac{1}{6}\frac{6\times 7\times 13}{6} - \frac{49}{4} = \frac{35}{12} \approx 2{,}92.$$

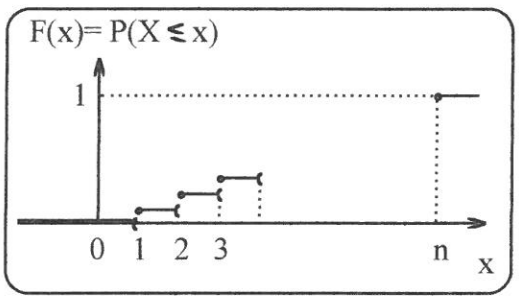
• *Remarque* :

Plus généralement, si X suit la loi uniforme U_n et que $X(\Omega) = [\![1, n]\!]$,

alors : $\forall\, k \in [\![1, n]\!]$, $P(X=k) = \frac{1}{n}$, $E(X) = \frac{n+1}{2}$ et $V(X) = \frac{n^2-1}{12}$.



Loi U_n



Fonction de répartition de $X \hookrightarrow U_n$

3°. Loi de Bernoulli (Jakob Bernoulli 1654-1705).

La variable aléatoire X suit la loi de Bernoulli de paramètre p, ce que l'on note :

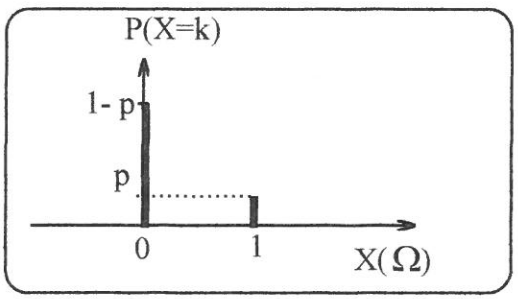
$X \hookrightarrow B(1, p)$, si : $X(\Omega) = \{0, 1\}$, $P(X=1) = p$ et $P(X=0) = 1 - p = q$.

Exemple **:**

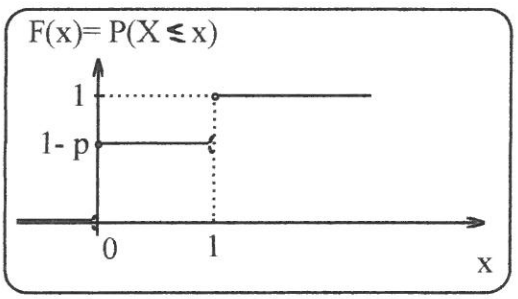
On lance un dé, la variable aléatoire X indique si l'événement "le dé donne un 6" est réalisé, c'est-à-dire que X est la fonction indicatrice de $\{6\}$.

$$P(X=1) = P(\{6\}) = \frac{1}{6} \text{ et } P(X=0) = P(\{1, 2, 3, 4, 5\}) = \frac{5}{6}.$$

$$E(X) = 1\times\frac{1}{6} + 0\times\frac{5}{6} = \frac{1}{6} \quad ; \quad V(X) = 1^2\times\frac{1}{6} + 0^2\times\frac{5}{6} - \left(\frac{1}{6}\right)^2 = \frac{5}{36}$$



Loi $B(1, p)$



Fonction de répartition de $X \hookrightarrow B(1, p)$

4°. Loi binomiale

La variable aléatoire X suit la loi binomiale de paramètres n et p, ce que l'on note :

$$X \hookrightarrow B(n, p), \text{ si : } X(\Omega) = [\![0, n]\!], \quad \forall k \in [\![0, n]\!], \ P(X = k) = \binom{k}{n} p^k (1-p)^{n-k}$$

On a alors : $E(X) = np$ et $V(X) = np(1-p)$.

Lien avec la loi de Bernoulli :

Soient $X_1, X_2, \ldots, X_n$ n variables aléatoires indépendantes suivant la loi de Bernoulli de paramètre p. La variable aléatoire $X = X_1 + X_2 + \ldots + X_n$ suit la loi $B(n, p)$.

Le schéma théorique :

On observe une succession de n épreuves aléatoires indépendantes à deux issues : "le succès" est obtenu avec la probabilité constante p et "l'échec" avec la probabilité $(1 - p)$. La variable aléatoire X, qui mesure le nombre de succès au cours de ces n épreuves suit alors la loi $B(n, p)$.

Le modèle d'urne :

Dans une urne, il y a des boules blanches en proportion p. On effectue un ***tirage avec remise*** (ou tirage bernoullien) de n boules dans cette urne. La variable aléatoire X qui mesure le nombre de boules blanches obtenu au cours de ces n épreuves suit la loi $B(n, p)$. (On admet que les épreuves, dans un tirage avec remise, sont indépendantes).

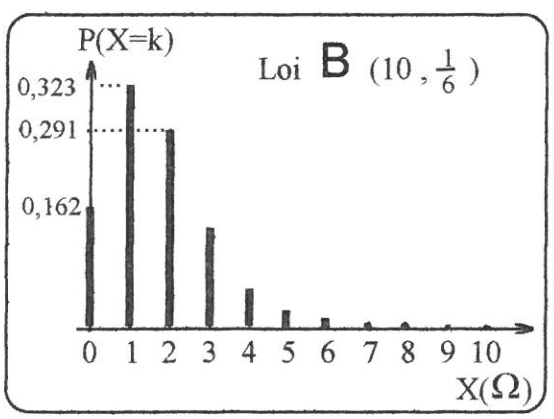
Précision orthographique : "binomiale" s'écrit sans accent circonflexe, comme "polynomial", mais "binôme" et "polynôme" en prennent un.

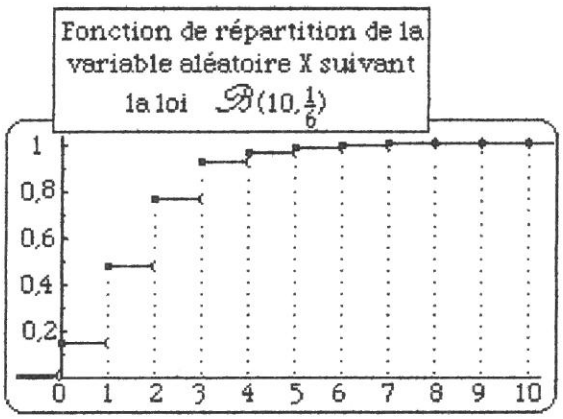
Remarque :

Le qualificatif "binomiale" vient de ce que lorsque l'on vérifie que les événements "X= k" pour $k \in [\![0, n]\!]$ forment un système complet d'événements, et donc que l'on est bien en présence d'une loi de probabilité, on est amené à utiliser la formule du binôme de Newton.

Exemple :

On lance un dé 10 fois, la variable aléatoire X qui mesure le nombre de fois où l'événement : "le dé donne un 6" est réalisé, au cours de ces 10 épreuves indépendantes, suit alors la loi $B(10, 1/6)$.





5°. Loi hypergéométrique

La variable aléatoire X suit la loi hypergéométrique de paramètres N, n et p, noté :

$$X \hookrightarrow H(N, n, p), \text{ si : } X(\Omega) = [\![\text{Max}(0, n - N + Np) ; \text{Min}(n, Np)]\!],$$

$$\text{et } \forall k \in X(\Omega),\ P(X = k) = \frac{C_{Np}^{k} \times C_{N(1-p)}^{n-k}}{C_{N}^{n}}$$

On a alors : $E(X) = np$ et $V(X) = np(1-p)\dfrac{N-n}{N-1}$.

Le modèle d'urne :

Dans une urne, il y a N boules dont des boules blanches en proportion p (donc Np boules blanches et par suite, $N(1-p)$ autres boules). On effectue un ***tirage sans remise*** ou ***exhaustif*** (ou encore "simultané") de n boules dans cette urne. La variable aléatoire X qui mesure le nombre de boules blanches obtenu parmi ces n boules suit la loi $H(N, n, p)$.

La taille de la population mère est N, la taille de l'échantillon est n, on cherche la probabilité d'obtenir k boules blanches exactement, donc n - k autres boules.

• Le nombre entier k vérifie les conditions :

$$\begin{cases} 0 \le k \le n \\ 0 \le k \le Np \end{cases} \text{ et } \begin{cases} 0 \le n-k \le n \\ 0 \le n-k \le N(1-p) \end{cases} \text{ d'où } k \in [\![\text{Max}(0, n - N + Np) ; \text{Min}(n, Np)]\!]$$

Remarque : Il suffit dans la pratique de savoir que $X(\Omega)$ est inclus dans $[\![1, n]\!]$.

• Le nombre de tirages possibles de n boules est C_{N}^{n}, et le nombre de ces tirages réalisant l'événement "on obtient k boules blanches et n - k autres boules" est $C_{Np}^{k} \times C_{N(1-p)}^{n-k}$,

d'où $P(X = k) = \dfrac{C_{Np}^{k} \times C_{N(1-p)}^{n-k}}{C_{N}^{n}}$.

Exemple :

Une classe de T.S. tertiaire est constituée de 15 garçons et de 20 filles. Un groupe de 18 élèves est désigné au hasard pour constituer un groupe de travaux pratiques. Soit X la variable aléatoire mesurant le nombre de filles de ce groupe.

On a : $X(\Omega) = [\![3, 18]\!]$, $X \hookrightarrow H(35, 18, 4/7)$.

• *Plus généralement, lorsqu'on effectue un prélèvement exhaustif dans une population, pour observer une sous-population, dont on connaît la proportion dans la population totale, le modèle probabiliste associé est celui d'une loi hypergéométrique.*

Par exemple, on sait que la production journalière d'un type de piles contient en moyenne 2 % d'objets défectueux, fréquence observée après une longue étude statistique, le critère étant que la durée de vie soit au moins de 200 heures pour que la pile soit déclarée non défectueuse.

On prélève un échantillon de 30 piles dans un lot de 1 000 dont on mesure la durée de vie : il s'agit d'un test destructif et donc d'un prélèvement sans remise. La variable aléatoire mesurant le nombre de piles défectueuses par échantillon exhaustif de taille 30 suit la loi H(1 000 ; 30 ; 0,02).

Au passage, on assimile la fréquence observée à une probabilité théorique.

Sous certaines conditions, la loi hypergéométrique peut être approchée par une loi binomiale, ce qui est l'objet du dernier paragraphe de cette étude.

6°. Loi géométrique

La variable aléatoire X suit la loi géométrique de paramètre p, notée $X \hookrightarrow \mathcal{G}(p)$,
si : $X(\Omega) = \mathbb{N}^*$, $\forall\, k \in \mathbb{N}^*$, $P(X = k) = p(1-p)^{k-1}$

On a alors : $E(X) = \dfrac{1}{p}$ et $V(X) = \dfrac{1-p}{p^2}$

Le modèle d'urne :

Dans une urne, il y a N boules dont une proportion p de boules blanches (donc Np boules blanches et par suite, $N(1-p)$ autres boules). On effectue des ***tirages avec remise*** dans cette urne.

La variable aléatoire X qui mesure le nombre des tirages jusqu'à l'obtention d'une boule blanche suit la loi $\mathcal{G}(p)$. C'est une loi du " temps d'attente ".

Exemple :

On lance un dé équilibré, la variable aléatoire X, qui mesure le nombre de jets jusqu'à ce que l'événement "le dé donne un as" soit réalisé, suit la loi $\mathcal{G}\left(\frac{1}{6}\right)$. $X(\Omega) = \mathbb{N}^*$.

Soit $k \in \mathbb{N}^*$, on a vu en préliminaire que $P(X = k) = \frac{1}{6}\left(\frac{5}{6}\right)^{k-1}$; $E(X) = 6$ et $V(X) = 30$.

Récréation :

Au cours d'un jeu télévisé, un candidat doit ouvrir une porte. Pour cela, il dispose d'une boîte qui contient dix clefs, d'allures semblables, mais dont une seule ouvre la porte.

Certains candidats, les "rationnels", utilisent la méthode qui consiste à essayer chaque clef prise au hasard dans la boîte, en y effectuant des tirages sans remise. D'autres candidats, les "désordonnés", essaient chaque clef prise au hasard dans la boîte, en y effectuant des tirages avec remise. Soient X et Y les variables aléatoires mesurant respectivement le nombre des clefs essayées pour obtenir l'ouverture de la porte par chacune des deux méthodes. Déterminer les lois de X et

de Y, leur espérance mathématique et leur écart type. Calculer la probabilité d'essayer plus de huit clefs par chacune des deux méthodes. On estime qu'un candidat sur trois utilise la méthode "désordonnée". Calculer la probabilité que le candidat soit classé parmi les désordonnés, sachant que les huit premiers essais ont échoué.

- Dans le cas du candidat rationnel, on a $X(\Omega) = \{1, 2, \ldots, 10\}$.

  Soit $k \in \{1, 2, \ldots, 10\}$ et A_k l'événement "la clef du $k^{ième}$ essai ouvre la porte".

  On a successivement : $P(X = 1) = P(A_1) = \frac{1}{10}$

  $$P(X = 2) = P(\overline{A_1} \cap A_2) = P_{\overline{A_1}}(A_2) \times P(\overline{A_1}) = \frac{1}{9} \times \frac{9}{10} = \frac{1}{10}.$$

  Par récurrence on obtient :

  $P(X = k) = P(\overline{A_1} \cap \overline{A_2} \cap \ldots \cap \overline{A_{k-1}} \cap A_k)$

  $$= P_{\overline{A_1} \cap \overline{A_2} \cap \ldots \cap \overline{A_{k-1}}}(A_k) \times P_{\overline{A_1} \cap \overline{A_2} \cap \ldots \cap \overline{A_{k-2}}}(A_{k-1}) \times \ldots \times P_{\overline{A_1}}(A_2) \times P(\overline{A_1})$$

  $$= \frac{1}{10 - (k - 1)} \times \frac{10 - (k - 1)}{10 - (k - 2)} \times \cdots \times \frac{8}{9} \times \frac{9}{10} = \frac{1}{10}.$$

  La variable aléatoire X suit donc la loi uniforme U_{10} d'où : $E(X) = \frac{11}{2}$ et $\sigma(X) = \sqrt{\frac{99}{12}} \approx 2{,}87$.

  Ainsi $P(X > 8) = P(X = 9) + P(X = 10) = 0{,}2$.

- Dans le cas du candidat désordonné, la variable aléatoire Y suit la loi géométrique de paramètre $\frac{1}{10}$. On a $Y(\Omega) = \mathbb{N}$, et pour tout $k \in \mathbb{N}^*$, $P(Y = k) = 0{,}1 \times (0{,}9)^{k-1}$.

  D'où : $E(Y) = 10$ et $\sigma(Y) = \sqrt{90} \approx 9{,}49$.

  $$P(Y > 8) = 1 - P(Y \leq 8) = 1 - \sum_{k=1}^{8} 0{,}1 \times (0{,}9)^{k-1} = 1 - 0{,}1 \times \sum_{k=1}^{7} (0{,}9)^k = 1 - 0{,}1 \times \frac{1 - 0{,}9^8}{1 - 0{,}9} = 0{,}9^8.$$

- Soit N la variable aléatoire mesurant le nombre d'essais nécessaires pour ouvrir la porte. Les événements $N = X$ et $N = Y$ forment un système complet et l'événement "le candidat est classé dans les désordonnés" est $N = Y$. Le théorème de Bayes donne :

$$P_{[N > 8]}(N = Y) = \frac{P_{[N = Y]}(N > 8) \times P(N = Y)}{P_{[N = Y]}(N > 8) \times P(N = Y) + P_{[N = X]}(N > 8) \times P(N = X)}$$

$$= \frac{P(Y > 8) \times \frac{1}{3}}{P(Y > 8) \times \frac{1}{3} + P(X > 8) \times \frac{2}{3}} = \frac{0{,}9^8}{0{,}9^8 + 0{,}2 \times 2} \approx 0{,}518.$$

7°. Loi de Poisson (Denis Poisson 1781-1840).

La variable aléatoire X suit la loi de Poisson de paramètre l, ce que l'on note $X \hookrightarrow \mathcal{P}(\lambda)$, si :

$$X(\Omega) = \mathbb{N},\ \forall k \in \mathbb{N},\ P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

On a alors : $E(X) = \lambda$ et $V(X) = \lambda$

Champ d'intervention : La loi de Poisson intervient dans la modélisation de phénomènes aléatoires "rares" où le futur est indépendant du passé tels que dans l'observation de :

- files d'attente par intervalles de temps assez courts (appels téléphoniques à un standard, arrivées de clients à un guichet),
- pannes de machines,
- sinistres enregistrés par une compagnie d'assurance pour une cause déterminée...

Obtention : On observe un phénomène dont l'apparition vérifie les conditions :

{ Indépendance des apparitions.
La probabilité d'une apparition entre les instants t et $t + \Delta t$ est proportionnelle à Δt, soit égale à $\lambda\, \Delta t$.
La probabilité qu'il y ait plus d'une apparition entre les instants t et $t + \Delta t$ est $o(\Delta t)$

Soit $P_n(t)$ la probabilité qu'il y ait n apparitions entre les instants 0 et t.

◆ Calcul de $P_0(t)$:

$P_0(t + \Delta t) = P(\text{"0 jusqu'à } t\text{"} \cap \text{"0 entre } t \text{ et } t + \Delta t\text{"}) = P(\text{"0 jusqu'à } t\text{"}) \times P(\text{"0 entre } t \text{ et } t + \Delta t\text{"})$

$= P_0(t) \times [1 - P(\text{"1 entre } t \text{ et } t + \Delta t\text{"}) - P(\text{"au moins 2 entre } t \text{ et } t + \Delta t\text{"})]$

$= P_0(t) \times [1 - \lambda \Delta t - o(\Delta t)]$

D'où $$\frac{P_0(t + \Delta t) - P_0(t)}{\Delta t} = -\lambda\, P_0(t) - P_0(t) \times \varepsilon(\Delta t)$$

Lorsque $\Delta t \to 0$, on a donc $P_0'(t) = -\lambda\, P_0(t)$ et par suite $P_0(t) = k\, e^{-\lambda t}$, la condition initiale $P_0(0) = 1$ donne $k = 1$, d'où $P_0(t) = e^{-\lambda t}$.

◆ Formule de récurrence sur n :

$P_n(t + \Delta t) = P(\text{"n jusqu'à } t\text{"} \cap \text{"0 entre } t \text{ et } t+\Delta t\text{"})$

$+ P(\text{"n - 1 jusqu'à } t\text{"} \cap \text{"1 entre } t \text{ et } t+\Delta t\text{"}) + P(\text{"n - 2 jusqu'à } t\text{"} \cap \text{"2 entre } t \text{ et } t+\Delta t\text{"})$

$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots + P(\text{"0 jusqu'à } t\text{"} \cap \text{"n entre } t \text{ et } t + \Delta t\text{"})$

$P_n(t + \Delta t) = P(\text{"n jusqu'à } t\text{"}) \times P(\text{"0 entre } t \text{ et } t+\Delta t\text{"})$

$+ P(\text{"n-1 jusqu'à } t\text{"}) \times P(\text{"1 entre } t \text{ et } t+\Delta t\text{"}) + P(\text{"n-2 jusqu'à } t\text{"}) \times P(\text{"2 entre } t \text{ et } t+\Delta t\text{"})$

$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots + P(\text{"0 jusqu'à } t\text{"}) \times P(\text{"n entre } t \text{ et } t + \Delta t\text{"})$

$$P_n(t + \Delta t) = P_n(t) \times [1 - \lambda \Delta t - o(\Delta t)] + P_{n-1}(t) \times \lambda \Delta t + P_{n-2}(t) \times o(\Delta t) + \ldots + P_0(t) \times o(\Delta t)$$

D'où $$\frac{P_n(t + \Delta t) - P_n(t)}{\Delta t} = -\lambda P_n(t) + \lambda P_{n-1}(t) + \varepsilon(\Delta t)$$

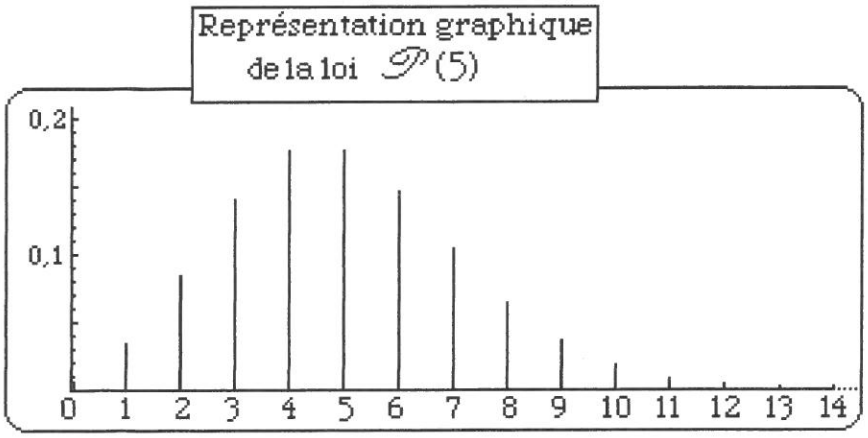
Lorsque $\Delta t \to 0$, on a donc $P_n'(t) = -\lambda P_n(t) + \lambda P_{n-1}(t)$.

- On montre par récurrence sur n que $P_n(t) = \frac{(\lambda t)^n e^{-\lambda t}}{n!}$, loi de Poisson $P(\lambda t)$.

Exemple **:**

Le nombre des clients se présentant au guichet "affranchissements" d'un bureau de poste par intervalle de temps de durée 10 minutes entre 14 heures et 18 heures, est mesuré par une variable aléatoire suivant une loi de Poisson de paramètre 5. La table du formulaire donne pour $\lambda = 5$, les probabilités des événements $P(X = k)$ pour tout entier k, $0 \leq k \leq 14$.

$k \backslash \lambda$		5
0		0,007
1		0,034
2		0,084
3		0,140
4		0,176
5		0,176
6		0,146
7		0,104
8		0,065
9		0,036
10		0,018
11		0,008
12		0,003
13		0,001
14		0,000



Calcul de la probabilité qu'entre 16 et 16 h 10 min. au moins 8 personnes se présentent au guichet :

$$P(X \geq 8) = 1 - P(X < 8) = 1 - [P(X = 0) + P(X = 1) + \ldots + P(X = 7)] = 1 - 0,867 = 0,133.$$

Calcul de la probabilité qu'entre 17 h 50 min. et 18 h, il y ait moins de 10 personnes à se présenter à ce guichet sachant qu'il y en a au moins 4 : $P_{(X \geq 4)}(X < 10) = \frac{P([X \geq 4] \cap [X < 10])}{P[X \geq 4]}) = \frac{P(4 \leq X \leq 9)}{1 - P(X < 4)}$

D'où $$P_{(X \geq 4)}(X < 10) = \frac{P(X = 4) + P(X = 5) + \ldots + P(X = 9)}{1 - [P(X = 0) + P(X = 1) + \ldots + P(X = 3)]} 0,956.$$

Nombre moyen d'usagers de ce guichet par heure au cours des tranches horaires journalières 14 h - 18 h, en supposant que les arrivées des clients soient des épreuves indépendantes :

l'espérance mathématique de la loi $P(5)$ est 5, on en déduit qu'en moyenne, cinq personnes se présentent au guichet par intervalle de dix minutes, la moyenne horaire du nombre des usagers de ce guichet de poste est 30.

❖ ❖ ❖

II. Lois continues

1°. Introduction

Comme on l'a vu dans le préliminaire, la **loi de probabilité** (ou *distribution*) d'une variable aléatoire continue X est caractérisée par sa fonction de répartition F ou par sa fonction densité de probabilité f.

L'espérance mathématique de X est, lorsqu'elle existe, le nombre réel

$$E(X) = \int_{-\infty}^{+\infty} t\, f(t)\, dt$$

La variance de X est, lorsqu'elle existe, le nombre réel $V(X) = E(X - E(X)]^2)$.

Théorème de König-Huygens : $V(X) = E(X^2) - [E(X)]^2 = \int_{-\infty}^{+\infty} t^2\, f(t)\, dt - [E(X)]^2$.

L'écart type de X est la racine carrée de la variance : $\sigma(X) = \sqrt{V(X)}$.

2°. Loi uniforme continue

La variable aléatoire X suit la loi uniforme U([a, b]), ce que l'on note : $X \hookrightarrow U([a, b])$, signifie : $X(\Omega) = [a, b]$ et la densité de probabilité de X est définie par :

$$\text{quel que soit } x \in [a, b],\ f(x) = \frac{1}{b - a} \quad \text{et} \quad \text{si } x \notin [a, b],\ f(x) = 0.$$

On a alors : $E(X) = \dfrac{a + b}{2}$ et $V(X) = \dfrac{(b - a)^2}{12}$.

***Exemple* :**

A un arrêt d'autobus, il passe un bus toutes les quinze minutes. Soit T la variable aléatoire mesurant le temps d'attente, exprimé en minutes, d'un usager arrivant à cette station. On suppose que la probabilité que l'attente soit inférieure ou égale à un temps t, pour $t \in [0,15]$, est proportionnelle à ce temps t. Quelle est la loi suivie par T ? Donner alors son espérance et son écart type. Calculer la probabilité qu'un usager pris au hasard arrivant à cette station attende plus de 10 minutes.

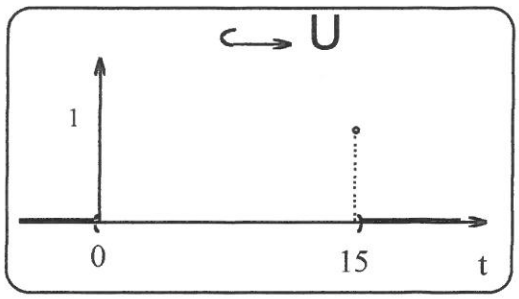
- Il existe un réel α, tel que pour $t \in [0,15]$, $P(T \leq t) = \alpha\, t$. Puisqu'il passe un bus toutes les quinze minutes à cette station, la probabilité d'attendre au plus 15 minutes est égale à 1, d'où $15\, \alpha = 1$. On en déduit que la fonction de répartition de la variable aléatoire T est définie par :

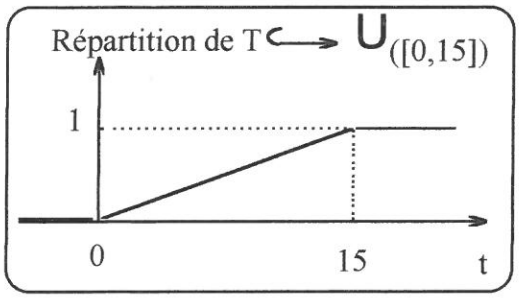
$$\begin{cases} F(t) = 0 & \text{si } t < 0 \\ F(t) = \frac{1}{15} t & \text{si } t \in [0, 15] \\ F(t) = 1 & \text{si } t > 1 \end{cases}$$

On en déduit que $T \hookrightarrow U([0,15])$.

- Le temps d'attente moyen est $E(T) = 7{,}5$ min. L'écart type est $\sigma(X) = \sqrt{18{,}75} \approx 4{,}33$ min.

• $P(T > 10) = 1 - P(T \leq 10) = 1/3$.





3°. Loi exponentielle

La variable aléatoire X suit la loi exponentielle de paramètre λ, noté $X \hookrightarrow E(\lambda)$, si :

$X(\Omega) = \mathbb{R}^+$ et que la densité de probabilité de X est définie par :

$$\forall x \in \mathbb{R}^-, f(x) = 0 \; ; \; \exists \lambda \in \mathbb{R}^+, \forall x \in \mathbb{R}^+, \; f(x) = \lambda e^{-\lambda x}.$$

On a alors : $E(X) = \frac{1}{\lambda}$ et $V(X) = \frac{1}{\lambda^2}$.

***Champ d'intervention* :**

La loi exponentielle intervient dans la modélisation de phénomènes aléatoires utilisés en fiabilité : loi de survie d'un équipement par exemple.

Des exemples ont été donnés : page 3, exemple n° 1 de variable aléatoire à densité.

page 5, exemple n° 4 .

4°. Loi normale (ou de Laplace-Gauss)

Pierre Simon de Laplace Mathématicien français (1749-1827)

Carl Friedrich Gauss Mathématicien allemand (1777-1855)

La variable aléatoire X suit la loi normale de paramètres m et σ, noté $X \hookrightarrow N(m, \sigma)$

lorsque : $X(\Omega) = \mathbb{R}$ et que la densité de probabilité de X est définie par :

$$\forall x \in \mathbb{R}, \; f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2}$$

On a alors : $E(X) = m$ et $V(X) = \sigma^2$.

***Champ d'intervention* :**

Dans la pratique, c'est la loi que l'on rencontre le plus souvent. Dès qu'un phénomène devient compliqué... il suit une loi normale ! (dans le secteur commercial, par exemple, la fluctuation des ventes autour de la moyenne dépend de nombreux paramètres, dans l'industrie, les diamètres de pièces usinées sont la résultante de la qualité de la matière première, du réglage de la machine, de l'usure de l'outil, de la température, etc...)

Loi normale centrée réduite :

Soit X une variable aléatoire continue suivant une loi normale de moyenne m et d'écart type σ. La variable aléatoire X - m est dite centrée car elle suit une loi normale de moyenne nulle et la variable aléatoire $T = \frac{X - m}{\sigma}$ est dite *centrée réduite* ou encore "normée" car elle suit une loi normale de moyenne 0 et d'écart type 1, notée N(0,1). La fonction densité de probabilité f de la variable aléatoire T suivant la loi N(0,1) est définie pour tout nombre réel x par :

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

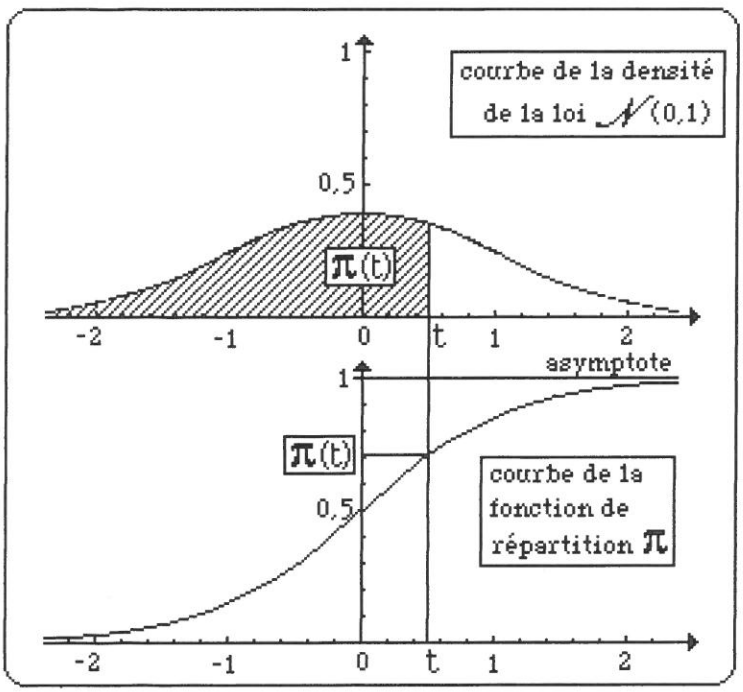
par suite, sa fonction de répartition est définie pour tout nombre réel t par :

$$\pi(t) = P(T \leq t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\frac{x^2}{2}} dx$$

La courbe de la fonction densité de probabilité d'une variable aléatoire suivant une loi normale est appelée courbe de Gauss ou aussi, de manière imagée, "courbe en cloche".

Propriétés :

Soit T une variable aléatoire continue suivant la loi N (0,1).



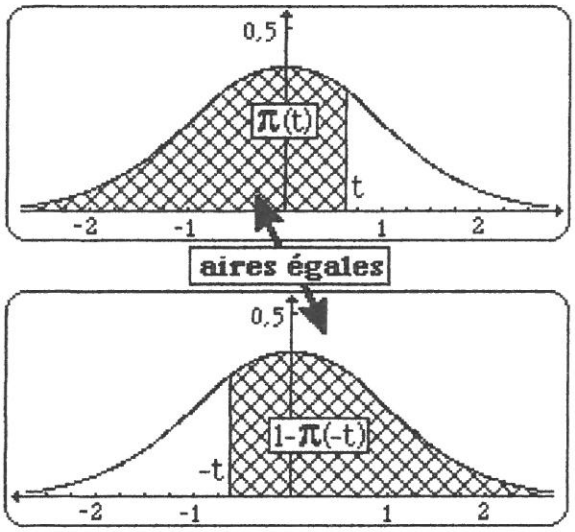
Toutes les relations suivantes sont établies pour un nombre réel t quelconque.

- Propriétés relatives à la ***continuité de la variable*** T :

$$\pi(T = t) = 0 \quad ; \quad \pi(T < t) = \pi(T \leq t)$$

- Propriété relative à la ***parité de la fonction densité*** f :

$$\pi(t) = 1 - \pi(-t)$$



La fonction densité de probabilité f de la variable aléatoire T est paire, la courbe représentative de f est symétrique par rapport à l'axe des ordonnées, c'est cette particularité qui est exploitée par la suite dans les calculs. En particulier, les aires mesurant les surfaces représentées ci-dessus sont égales et en tenant compte de ce que *l'aire de la surface totale située "sous la courbe de Gauss" est égale à 1*, on obtient la relation énoncée.

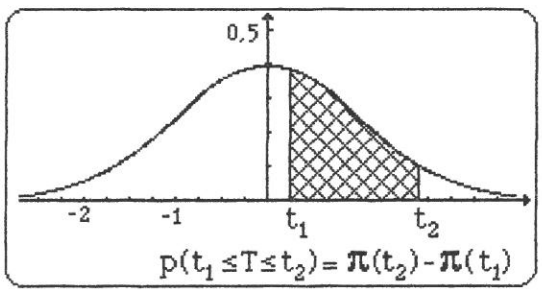
On observe, pour t=0, que $\pi(0) = \frac{1}{2}$.

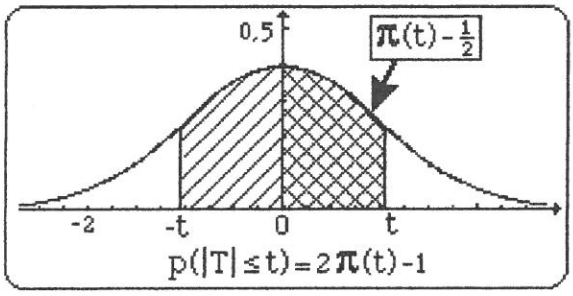
- Exemple de calcul de $P(t_1 \leq T \leq t_2)$ avec $t_1 = 0{,}25$ et $t_2 = 1{,}65$:

$$P(0{,}25 \leq T \leq 1{,}65) = \pi(1{,}65) - \pi(0{,}25) = 0{,}9495 - 0{,}5987 = 0{,}3508$$

Cas particulier d'un intervalle centré autour de la moyenne 0 : calcul de $P(|T| < 1{,}05)$.

$$P(|T| < 1{,}05) = 2\,P(0 < T < 1{,}05) = 2\,(\pi(1{,}05) - \pi(0)) = 2\,\pi(1{,}05) - 1 \approx 0{,}7062.$$





C. Approximations

I. Tirage bernoullien et tirage exhaustif :

1°. Tirage bernoullien :

Une urne contient vingt boules blanches et trente boules noires, les boules sont indiscernables au toucher. (*Lorsqu'on suppose les boules indiscernables au toucher, il faut comprendre que l'on se place délibérément dans une situation d'équiprobabilité, c'est-à-dire que lors d'un tirage, chaque boule a la même probabilité d'être prélevée.*)

On prélève successivement quatre boules, les tirages sont effectués avec remise, c'est-à-dire qu'après avoir noté la couleur de la boule, celle-ci est remise dans l'urne. (*Lorsque des tirages sont effectués avec remise, ces épreuves sont indépendantes ainsi que les événements liés à ces épreuves.*) On a une succession de quatre épreuves indépendantes ayant deux issues de probabilités constantes.

La variable aléatoire X mesurant le nombre de boules blanches obtenues au cours des quatre tirages suit alors la loi binomiale B(4 ; 0,4). On en déduit :

k	0	1	2	3	4
P(X = k)	0,1296	0,3456	0,3456	0,1536	0,0256

$E(X) = 1,6$; $V(X) = 0,96$; $\sigma_X \approx 0,98$.

2°. Tirage exhaustif :

On garde la même urne contenant vingt boules blanches et trente boules noires indiscernables au toucher. On prélève alors *simultanément* quatre boules. Soit Y la variable aléatoire donnant le nombre de boules blanches obtenues. Déterminons la loi de Y, on sait en fait que Y suit la loi $H(50 ; 4 ; 2/5)$. On a $Y(\Omega)=\{0, 1, 2, 3, 4\}$.

Soit k un élément de $Y(\Omega)$. Le nombre de combinaisons de quatre boules parmi cinquante est C_{50}^{4}, c'est le nombre des cas possibles. Le nombre des combinaisons de k boules blanches parmi vingt est C_{20}^{k}, chacune de ces combinaison est complétée par un ensemble de 4 - k boules noires. Le nombre des combinaisons de 4 - k boules noires parmi trente est C_{30}^{4-k} on en déduit que le nombre des cas favorables à la réalisation de l'événement Y= k est $C_{20}^{k} \times C_{30}^{4-k}$.

La loi de Y est alors définie par : pour tout k de {0,1,2,3,4}, $P(Y = k) = \dfrac{C_{20}^{k} \times C_{30}^{4-k}}{C_{50}^{4}}$

On en déduit :

k	0	1	2	3	4
P(Y = k)	0,119	0,3526	0,3589	0,1485	0,021

$E(Y) = 1,6$; $V(Y) \approx 0,9012$; $\sigma_Y \approx 0,95$.

On observe que la loi de Y est "proche" de celle de X et que $E(X) = E(Y)$.

Soit une population de taille N constituée de deux types d'individus, en nombre respectivement N_1 et N_2. On admet usuellement que la loi de la variable aléatoire Y donnant le nombre d'individus du premier type obtenus par tirage exhaustif d'un échantillon de taille n dans cette population de taille N, peut être approchée par la loi binomiale B(n, p) lorsque $N \geq 10n$.

Cette condition d'approximation n'a pas à être mémorisée. Lorsqu'il est décidé d'assimiler un tirage exhaustif à un tirage bernoullien, l'indication doit être donnée dans le texte de l'exercice et la justification n'est pas exigible.

II. Approximation d'une loi binomiale par une loi de Poisson:

Dans une entreprise, une étude statistique a montré qu'en moyenne 5% des articles d'une chaîne de fabrication présentent des défauts. Lors d'un contrôle de qualité, on envisage de prélever un échantillon de 120 articles. Bien que ce prélèvement soit exhaustif, on considère que la production est suffisamment nombreuse pour que l'on puisse assimiler cette épreuve à un tirage avec remise et que la probabilité qu'un article prélevé soit défectueux est constante. La variable aléatoire X donnant le nombre d'articles défectueux d'un tel échantillon suit alors la loi binomiale B(120 ; 0,05) et l'espérance mathématique de X est $120 \times 0,05 = 6$.

Comparons la loi de X avec celle d'une variable aléatoire Y suivant la loi de Poisson P(6) :

k	lois de X	lois de Y
0	0,002	0,002
1	0,013	0,015
2	0,042	0,045
3	0,087	0,089
4	0,134	0,134
5	0,163	0,161
6	0,165	0,161
7	0,141	0,138
8	0,105	0,103
9	0,069	0,069
10	0,040	0,041
11	0,021	0,023
12	0,010	0,011
13	0,004	0,005
14	0,002	0,002
15	0,001	0,001
16	0,000	0,000
...	.....	.....

On observe que la loi de la variable Y est suffisamment proche de celle de X pour que l'on puisse utiliser la loi de Poisson pour calculer par exemple la probabilité qu'un échantillon de 120 articles contienne au moins un article défectueux, puis la probabilité que cet échantillon contienne au plus trois articles défectueux. Présentation des calculs :

$P(X \geq 1) = 1 - P(X < 1) = 1 - P(Y < 1) = 1 - P(Y = 0) = 0,998.$

$P(X \leq 3) = P(Y \leq 3) = P(Y = 0) + P(Y = 1) + P(Y = 2) + P(Y = 3)$

d'où $P(X \leq 3) = 0,151.$

(On observe que l'erreur commise est de l'ordre de 7×10^{-3}.)

L'intérêt de cette approximation réside essentiellement dans le fait que la loi de Poisson ne dépend que d'un seul paramètre, alors que la loi binomiale dépend de deux paramètres.

On admet que sous certaines conditions, la loi d'une variable aléatoire binomiale B(n, p) peut être approchée par celle d'une variable aléatoire $\tilde{X}$ suivant une loi de Poisson P(λ) de paramètre $\lambda = n\,p$.

Des conditions usuelles de validité d'approximation d'une loi binomiale B(n, p) par une loi de Poisson sont : $n \geq 30$, $p \leq 0{,}1$ et $n\,p < 15$.

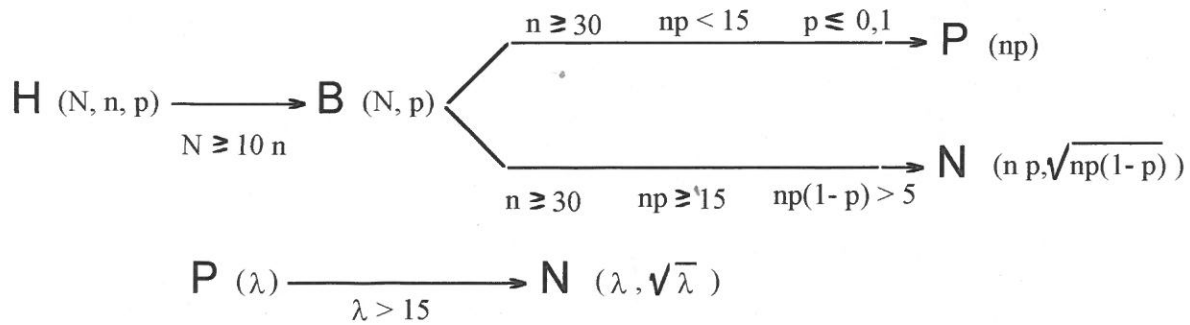
Les conditions d'approximation d'une loi binomiale par une loi de Poisson n'ont pas à être mémorisées. La seule capacité exigible est de savoir que lorsqu'une loi binomiale B(n, p) peut être approchée par une loi de Poisson P(λ), le paramètre de cette loi est donné par $\lambda = n\,p$, c'est-à-dire que les variables aléatoires X suivant la loi B(n, p) et $\tilde{X}$ dont la loi P(λ) est une approximation de la loi de X ont la même moyenne :

$$E(X) = E(\tilde{X}) = n\,p = \lambda.$$

Idées essentielles :

- Il y a **conservation de la moyenne** dans chaque approximation et éventuellement de l'écart type.
- Il convient d'appliquer la **correction de continuité** dans le cas de l'approximation d'une loi discrète par une loi continue.
- Aucune connaissance concernant les conditions d'approximation n'est exigible des élèves de S.T.S.

Conditions usuelles d'approximations



La correction de continuité

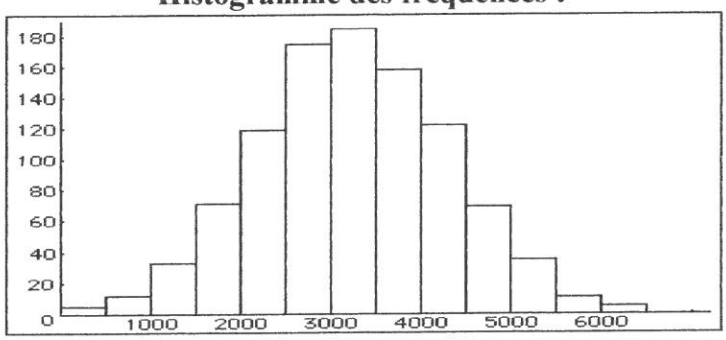
A. Histogramme et polygone des fréquences en statistique descriptive

Dans une banque, on a relevé pendant un mois les montants des retraits en espèces de mille clients. La population observée est l'ensemble des mille clients suivant le caractère "montant des retraits effectués en espèces au cours du mois". Il s'agit d'un caractère quantitatif discret ; mais lors du regroupement des données en classes, bien que les valeurs observées soient entières ou décimales d'ordre deux, la variable statistique associée sera traitée comme une variable continue pouvant donc prendre toute valeur d'un intervalle de réels. La variable statistique associée fait correspondre à chaque client la modalité retenue suivant le montant de ses retraits. La série statistique présentée associe à chaque modalité [0,500[, [500,1000[, ..., [6000, +∞[le nombre des clients dont les retraits sont situés dans cet intervalle. On a les résultats suivants :

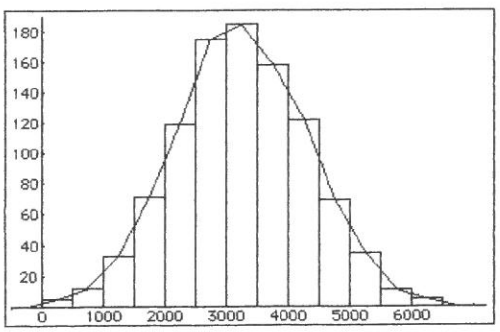
Montant des retraits exprimés en francs	Nombre de clients	Fréquence
moins de 500	5	0,005
de 500 à moins de 1000	12	0,012
de 1000 à moins de 1500	33	0,033
de 1500 à moins de 2000	71	0,071
de 2000 à moins de 2500	119	0,119
de 2500 à moins de 3000	175	0,175
de 3000 à moins de 3500	185	0,185
de 3500 à moins de 4000	158	0,158
de 4000 à moins de 4500	122	0,122
de 4500 à moins de 5000	69	0,069
de 5000 à moins de 5500	35	0,035
de 5500 à moins de 6000	11	0,011
6000 et plus	5	0,005
TOTAL	**1000**	**1**

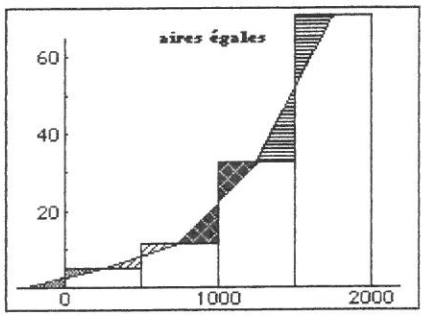
(source confidentielle)

Histogramme des fréquences :



Dans cet exemple, les douze premières classes ont la même amplitude, on représente conventionnellement la dernière classe [6000,+∞[avec cette même amplitude.





Le polygone des effectifs est obtenu en joignant les milieux des côtés supérieurs des rectangles de l'histogramme, avec la convention usuelle concernant les classes extrêmes pour que l'aire du domaine de frontières l'axe des abscisses et le polygone statistique soit égale à la somme des aires

des rectangles de l'histogramme. La courbe tracée traduit la *continuité* de la série statistique étudiée.

B. Propriété pratique de la loi $\mathcal{N}(m, \sigma)$

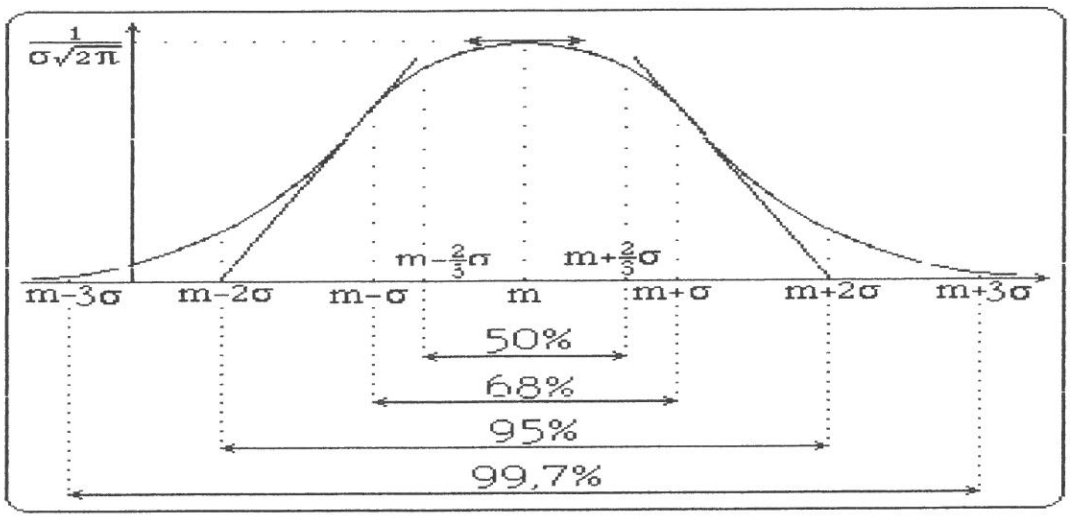
Soit X une variable aléatoire continue suivant une loi normale de moyenne m et d'écart type σ et T la variable aléatoire normale centrée réduite associée à X.

La relation $X = \sigma T + m$ permet d'affirmer que pour tout nombre réel strictement positif k, on a :

$$P(m - k\sigma < X < m + k\sigma) = P(-k < T < k) = 2\pi(k) - 1$$

et pour $k \in \{\frac{2}{3}, 1, 2, 3\}$, on obtient : $P(-\frac{2}{3} < T < \frac{2}{3}) = 0{,}4951$, $P(-1 < T < 1) = 0{,}6826$;

$$P(-2 < T < 2) = 0{,}9544 \ ; \ P(-3 < T < 3) = 0{,}9973.$$



On en déduit que lorsqu'un phénomène suit une loi normale,

50% des observations sont à moins de deux tiers d'écart type de la moyenne,

68% des observations sont à moins d'un écart type de la moyenne,

95% des observations sont à moins de deux écarts types de la moyenne,

99,7 % à moins de trois écarts types de la moyenne.

C. Approximation d'une loi discrète par une loi normale

II. Ajustement analytique

Le directeur d'une entreprise de vente de matériaux de construction fait établir une statistique sur les ventes mensuelles de ciment en sacs de 50 kilogrammes.

Les résultats sont consignés dans le tableau suivant :

Nombres de sacs	Pourcentages
[3700,4200[	5
[4200,4700[	19
[4700,5200[	35
[5200,5700[	29
[5700,6200[	10
[6200,6700[	2

Le calcul de la demande moyenne m et de l'écart type σ de cette série statistique a donné :

$$m = \sum_{i=1}^{6} c_i f_i = 5\,080 \quad \text{et} \quad \sigma = \sqrt{\left(\sum_{i=1}^{6} c_i f_i\right) - m^2} \approx 546.$$

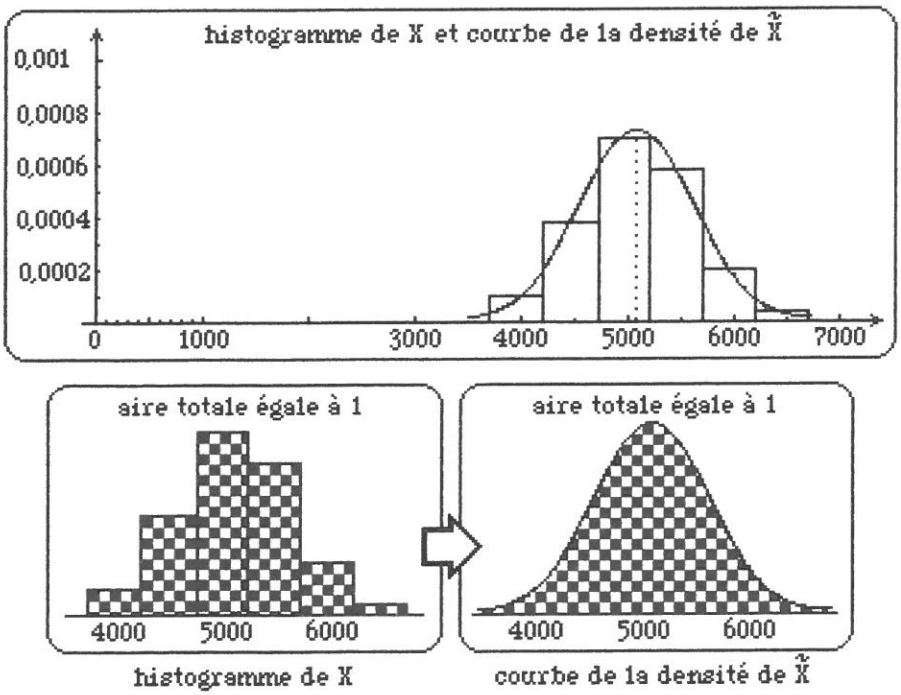
Les résultats du calcul des pourcentages des observations situées dans les intervalles, obtenus par interpolation affine, $[m - k\sigma ; m + k\sigma]$ où $k \in \{\frac{2}{3}, 1, 2, 3\}$, sont donnés ci-contre :

Nombres de sacs	Pourcentages observés	Pourcentages pour une loi normale
[4716,5444]	48,03	49,51
[4534,5626]	66,02	68,26
[3988,6172]	94,56	95,44
[3442,6718]	100	99,73

Les pourcentages observés étant suffisamment proches des pourcentages déterminés dans le paragraphe précédent pour une loi normale, on décide d'approcher la loi de la variable aléatoire X donnant le nombre des ventes mensuelles par la loi $\mathcal{N}(5080, 546)$.

Interprétation graphique

Soit $\tilde{X}$ une variable aléatoire **continue** suivant la loi normale $\mathcal{N}(5080, 546)$ et f sa fonction densité de probabilité.



L'approximation effectuée est représentée par le graphique ci-dessus : on remplace l'histogramme de la variable X construit de telle sorte que l'aire totale de l'histogramme soit égale à 1 par celui de $\widetilde{X}$.

Un problème se pose alors. La variable aléatoire X est discrète, elle prend toute valeur entière de 3700 à 6700, avec une probabilité non nulle, par exemple $P(X=5000) \neq 0$.

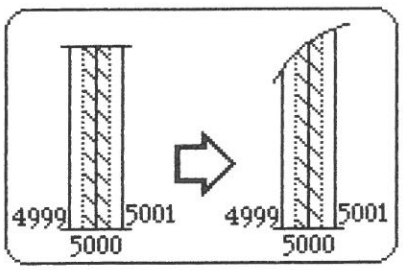
La variable $\widetilde{X}$ est continue, elle prend toute valeur réelle et la probabilité que $\widetilde{X}$ prenne une valeur donnée est nulle, par exemple $P(\widetilde{X} = 5000) = 0$.

On est donc amené à appliquer ce qu'on appelle ***la correction de continuité***, procédé décrit dans les exemples suivants.

1°. Puisque X ne prend que des valeurs entières, on a :

$$P(X = 5000) = P(4999{,}5 \leq X \leq 5000{,}5).$$

On remplace alors X par $\widetilde{X}$, ce qui revient à approcher l'aire du rectangle construit sur le segment représentant l'intervalle [4999,5 ; 5000,5] dans l'histogramme de X, par l'aire du domaine inscrit sous la courbe de la fonction densité de probabilité de $\widetilde{X}$ et sur le même intervalle, comme le montre le graphique ci-dessous. On poursuit alors le calcul.



Soit T la variable normale centrée réduite associée à $\widetilde{X}$. On a $T = \dfrac{\widetilde{X} - 5080}{546}$, d'où :

$$P(X = 5000) = P(4999{,}5 \leq X \leq 5000{,}5) = P(4999{,}5 \leq \widetilde{X} \leq 5000{,}5) = P(-0{,}1474 \leq T \leq -0{,}1456)$$

$$= P(0{,}1456 \leq T \leq 0{,}1474) = \pi(0{,}1474) - \pi(0{,}1456) \approx 0{,}0007.$$

2°. Calcul de la probabilité de l'événement :

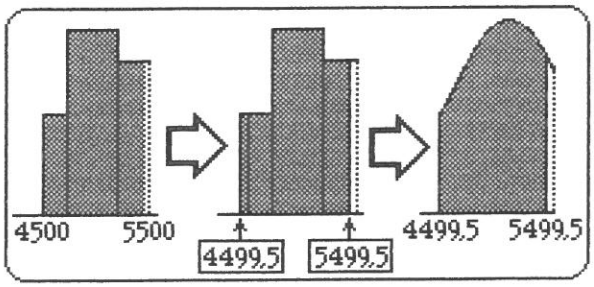
"le nombre des ventes est supérieur ou égal à 4500 et strictement inférieur à 5500".

Avec les notations définies dans l'exemple précédent, on a :

$$P(4500 \leq X < 5500) = P(4499{,}5 \leq X \leq 5499{,}5) = P(4499{,}5 \leq \widetilde{X} \leq 5499{,}5)$$

$$= P(-1{,}0632 \leq T \leq 0{,}7683) = \pi(0{,}7683) - \pi(-1{,}0632)$$

$$= \pi(0{,}7683) + \pi(1{,}0632) - 1 \approx 0{,}6349.$$

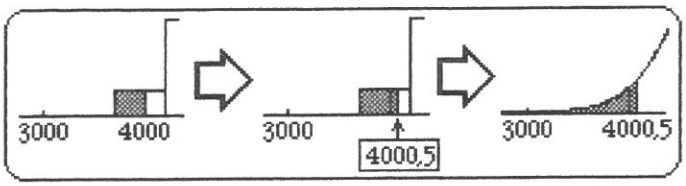


3°. Calcul de la probabilité de l'événement :

"le nombre des ventes est inférieur ou égal à 4000 ".

$$P(X \leq 4000) = P(X \leq 4000,5) = P(\tilde{X} \leq 4000,5) = P(T \leq -1,9771) = \pi(-1,9771)$$

$$= 1 - \pi(1,9771) \approx 0,024.$$



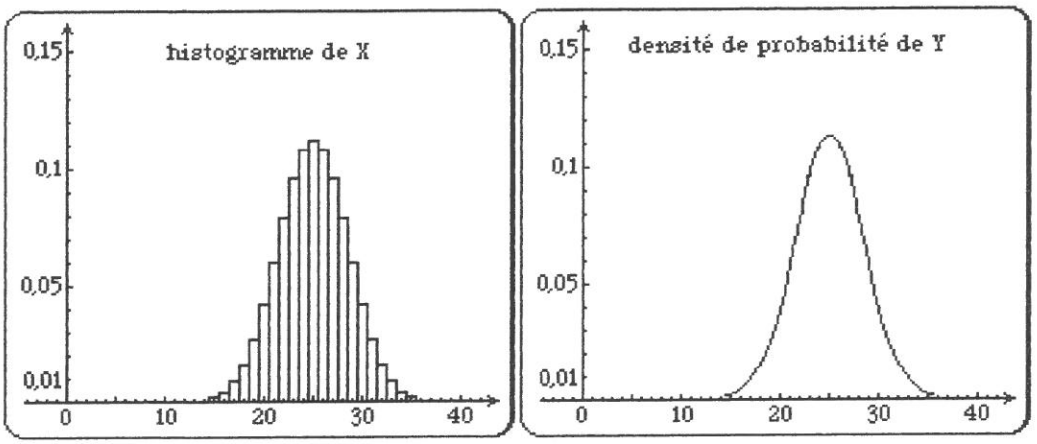
B. Cas d'une loi binomiale

On lance cinquante fois de suite une pièce de monnaie équilibrée.

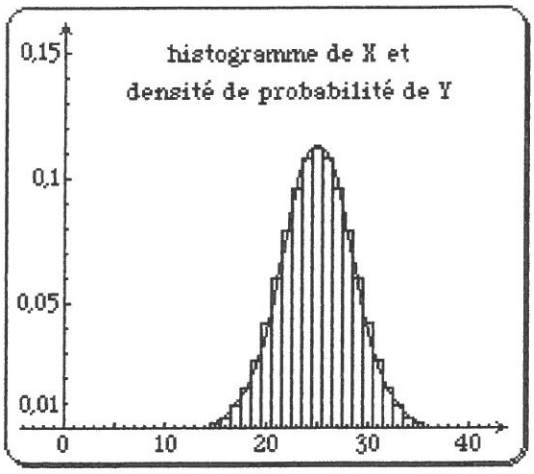
Soit X la variable aléatoire donnant le nombre de "face". On sait que X suit la loi binomiale $\mathcal{B}(50;0,5)$.

L'espérance mathématique de X est $E(X)=25$ et son écart type est $\sigma = 3,54$.

Traçons l'histogramme de X et la fonction densité de probabilité d'une variable aléatoire normale $\mathcal{N}(25 ; 3,54)$.



On observe que les histogrammes des variables aléatoires X et Y coïncident pratiquement.



On admet que sous certaines conditions, la loi d'une variable aléatoire binomiale $\mathcal{B}(n, p)$ peut être approchée par celle d'une variable aléatoire $\widetilde{X}$ suivant la loi normale $\mathcal{N}(n\,p, \sqrt{np\,(1-p)})$.

Des conditions usuelles de validité d'approximation d'une loi binomiale par une loi normale sont :

$$n \geq 30 \quad ; \quad n\,p \geq 15 \quad \text{et} \quad n\,p(1-p) > 5.$$

L'intérêt de cette approximation est de simplifier les calculs.

Dans les classes de Techniciens Supérieurs, les conditions d'approximation d'une loi binomiale par une loi normale n'ont pas à être mémorisées. La seule capacité exigible des élèves est de savoir que lorsqu'une loi binomiale $\mathcal{B}(n, p)$ peut être approchée par une loi normale $\mathcal{N}(m, \sigma)$, les paramètres de cette loi sont donnés par $m = n\,p$ et $\sigma = \sqrt{n\,p\,(1-p)}$, c'est-à-dire que les variables aléatoires X suivant la loi $\mathcal{B}(n, p)$ et $\widetilde{X}$ dont la loi $\mathcal{N}(m, \sigma)$ est une approximation de la loi de X ont la même moyenne et le même écart type : $E(X) = n\,p = E(\widetilde{X}) = m$ et $\sigma(X) = \sigma(\widetilde{X}) = \sqrt{n\,p\,(1-p)}$.

***Exemple* :**

Une étude sur le comportement des automobilistes a permis de constater que, dans le centre d'une grande ville, 10% des stationnements étaient irréguliers. Une contractuelle contrôle 800 véhicules par jour et dresse une contravention pour tout véhicule en stationnement irrégulier. Soit C la variable aléatoire donnant le nombre quotidien des contraventions dressées par cette contractuelle. On est en présence d'une succession de 800 épreuves indépendantes ayant deux issues avec une probabilité constante et la variable aléatoire C donne le nombre des succès obtenus. (On remarquera qu'il s'agit bien d'un modèle de tirage avec remise, un même véhicule pouvant bénéficier plusieurs fois au cours de la journée de l'attention de cette contractuelle.) On en déduit

que la variable aléatoire C suit la loi binomiale $\mathcal{B}(800 ; 0,1)$. L'espérance mathématique de C est 80 et son écart type est 8,49.

En utilisant l'approximation de la loi de C par une loi normale, calculer la probabilité de chacun des événements suivants :

E_1 "la contractuelle dresse moins de 50 contraventions",

E_2 "la contractuelle dresse plus de 90 contraventions",

E_3 "la contractuelle dresse moins de 100 contraventions sachant qu'elle en a dressé au moins 60".

◆ *Solution* :

Soit $\widetilde{C}$ une variable aléatoire continue suivant la loi normale $\mathcal{N}(80, 8,49)$ et T la variable normale centrée réduite associée à $\widetilde{C}$. On a alors :

$$P(E_1) = P(C < 50) = P(C \leq 49,5) = P(\widetilde{C} \leq 49,5) = P(T \leq -3,5925) = 1 - \pi(3,5925) \approx 0,000159.$$

$$P(E_2) = P(C > 90) = P(C \geq 90,5) = P(\widetilde{C} \geq 90,5) = P(T \geq 1,2367) = 1 - \pi(1,2367) \approx 0,1081.$$

$$P(E_3) = P_{[C \geq 60]}(C < 100) = \frac{P([C \geq 60] \cap [C < 100])}{P([C \geq 60])} = \frac{P(60 \leq C < 100)}{P([C \geq 60])} = \frac{P(59,5 \leq C \leq 99,5)}{P([C \geq 59,5])} = \frac{2}{3}$$

Ayant appliqué la correction de continuité, on remplace alors C par $\widetilde{C}$:

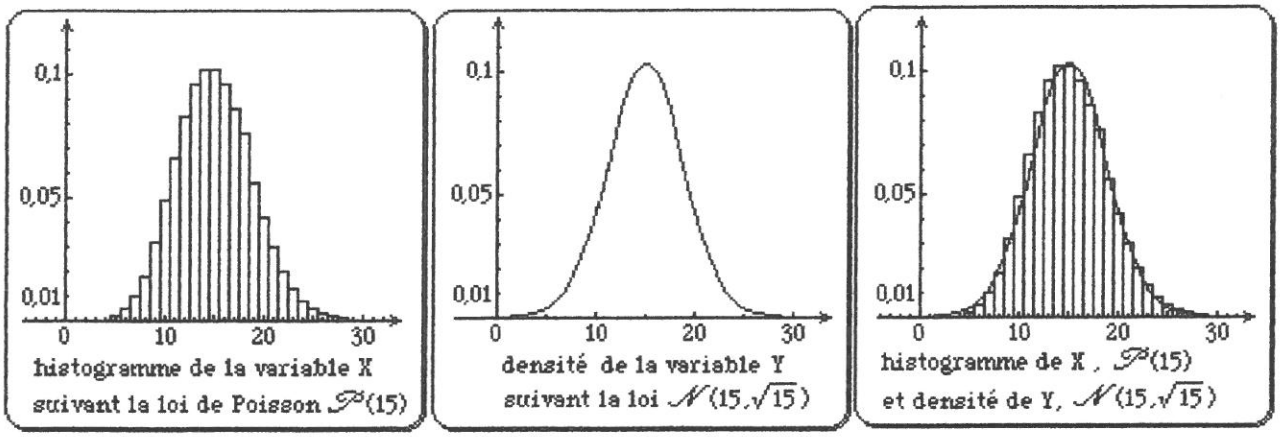
$$P(E_3) = \frac{P(59,5 \leq \widetilde{C} \leq 99,5)}{P([\widetilde{C} \geq 59,5])} = \frac{P(-2,4146 \leq T \leq 2,2968)}{P([T \geq -2,4146])} = \frac{\pi(2,2968) - (1 - \pi(2,4146))}{\pi(2,4146)} \approx 0,9891.$$

D. Cas d'une loi de Poisson

(Ce cas n'est pas au programme des classes de Techniciens Supérieurs.)

Dans la gare d'une grande ville le nombre des usagers utilisant l'un des composteurs de billets par quart d'heure de 6 heures 15 minutes à 21 heures du lundi au vendredi, définit une variable aléatoire X suivant une loi de Poisson de paramètre 15.

L'espérance mathématique de X est donc 15 et son écart type est $\sqrt{15} \approx 3,87$. Traçons l'histogramme de X et la fonction densité de probabilité d'une variable aléatoire normale $\mathcal{N}(15, 3,87)$.



On observe que les histogrammes des variables aléatoires X et Y coïncident pratiquement.

On admet que sous certaines conditions, la loi d'une variable aléatoire X suivant la loi de Poisson $\mathcal{P}(\lambda)$ peut être approchée par celle d'une variable aléatoire $\tilde{X}$ suivant la loi normale† $\mathcal{N}(\lambda, \sqrt{\lambda})$.

Une condition de validité d'approximation d'une loi de Poisson par une loi normale est $\lambda \geq 15$.

L'intérêt de cette approximation est de simplifier les calculs.

Lorsqu'une loi de Poisson $\mathcal{P}(\lambda)$ peut être approchée par une loi normale $\mathcal{N}(m, \sigma)$, les paramètres de cette loi sont donnés par $m = \lambda$ et $\sigma = \sqrt{\lambda}$, c'est-à-dire que les variables aléatoires X suivant la loi $\mathcal{P}(\lambda)$ et $\tilde{X}$ dont la loi $\mathcal{N}(m, \sigma)$ est une approximation de la loi de X ont la même moyenne et le même écart type : $E(X) = E(\tilde{X}) = \lambda$ et $\sigma(X) = \sigma(\tilde{X}) = \sqrt{\lambda}$.

Il convient également d'appliquer la correction de continuité.

***Exemple* :**

Dans la situation présentée, qui consiste à observer l'utilisation d'un composteur de billets dans une gare, déterminer la probabilité de chacun des événements suivants :

E_1 "en un quart d'heure, plus de vingt voyageurs utilisent ce composteur de billets"

E_2 "en un quart d'heure, au plus douze voyageurs utilisent ce composteur de billets"

E_3 "en un quart d'heure, moins de vingt voyageurs vont utiliser cet appareil sachant qu'il y en a au moins neuf."

◆ *Solution* :

On sait que X suit la loi de Poisson de paramètre 15. Soit $\tilde{X}$ une variable aléatoire suivant la loi normale $\mathcal{N}(15, \sqrt{15})$ et T la variable centrée réduite associée à $\tilde{X}$.

On obtient alors :

$$P(E_1) = P(X > 20) = P(X \geq 20{,}5) = P(\tilde{X} \geq 20{,}5) = P\left(T \geq \frac{5{,}5}{\sqrt{15}}\right) = 1 - \pi(1{,}42) \approx 0{,}0778.$$

$$P(E_2) = P(X \leq 12) = P(X \leq 12{,}5) = P(\tilde{X} \leq 12{,}5) = P\left(T \leq \frac{-2{,}5}{\sqrt{15}}\right) = 1 - \pi(1{,}118) \approx 0{,}1318.$$

$$P(E_3) = P_{[X \geq 9]}(X < 20) = \frac{P(9 \leq X < 20)}{P(X \geq 9)} = \frac{P(8{,}5 \leq X \leq 19{,}5)}{P(X \geq 8{,}5)} = \frac{P(8{,}5 \leq \tilde{X} \leq 19{,}5)}{P(\tilde{X} \geq 8{,}5)}$$

$$= \frac{P(-2{,}9069 \leq T \leq 2{,}0125)}{P(T \geq -2{,}9069)} = \frac{\pi(2{,}0125) - (1 - \pi(2{,}9069))}{\pi(2{,}9069)} \approx 0{,}978.$$

❖ ❖ ❖

Probabilités et Statistiques des problèmes

« Et si l'on observe ensuite que dans les choses qui peuvent ou non être soumises au calcul, la théorie des probabilités ... apprend à se garantir des illusions. Il n'est pas de science qu'il soit plus utile de faire entrer dans le système de l'instruction publique. »

(Laplace 1812)

Introduction

Dans l'enseignement des mathématiques en STS, pourquoi ne pas donner la priorité à un enseignement développant un apprentissage qui prenne appui sur des situations significatives et donc plus motivantes ?

Quels outils utiliser ou élaborer pour répondre aux questions que l'on se pose, en prise directe avec son environnement ? Mêler intimement problèmes, conjectures, simulations, compréhension de phénomènes « naturels », et résultats mathématiques, cela change de la démarche de construction magistrale d'une théorie suivie d'exercices d'application. Cela permet de sortir de l'algorithme « j'apprends, j'applique » cher à nos élèves.

D'une manière peu scolaire, Arthur Engel dans son ouvrage « Les certitudes du hasard, ALEAS Editeur » présente de nombreux thèmes d'étude évitant la traditionnelle dichotomie probabilités-statistiques.

Cet ouvrage dont la lecture est attrayante, présente les particularités suivantes :

- utilisation de vraies données statistiques tirées de publications scientifiques ou économiques.
- utilisation pratique des ordinateurs et machines programmables (la part théorique est ainsi réduite à l'essentiel). L'étude des différentes distributions statistiques se fait avec utilisation minimum des tables.
- les variables aléatoires réelles sont pour la plus grande part indépendantes (ou faiblement dépendantes).
- pour le théorème central limite, on se met dans le cas d'une variable distribuée quasi normalement sur l'intervalle [0,1]. Ce n'est que pour la standardisation que l'on a besoin de l'espérance, de la variance et d'un petit peu plus de théorie.

- l'usage des calculatrices permet d'introduire les intervalles de confiance à la place du manichéen -significatif -non significatif.

- Un travail plus théorique sur les probabilités conditionnelles n'est pas nécessaire d'entrée de jeu. C'est une notion difficile dont l'étude théorique peut-être placée tardivement.

La lecture avec papier-crayon de cet ouvrage apporte beaucoup à notre enseignement dans des sections où manque cruellement une « vision » d'archétypes servant à modéliser des situations diverses que les étudiants ont des difficultés à appréhender.

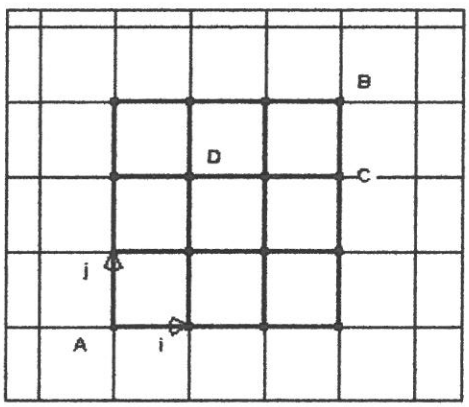
Illustrons le propos

Examinons ensemble quelques exemples tirés de l'ouvrage de Engel ou d'ouvrages scolaires et qui serviront à illustrer l'ensemble du propos, autour de trois « archétypes » pour lesquels une vision figurée est une aide précieuse :

1. Vision des C_n^p.

La connaissance archétypale des combinaisons est celle des parcours aléatoires sur une grille quadrillée où deux sens de déplacements sont autorisés.

□ **Exercice 1**



On dispose d'un quadrillage de repère $((A, \vec{i}, \vec{j})$ comme indiqué ci-contre. On appelle chemin de A à B toute succession de quatre déplacements élémentaires. Les seuls déplacements élémentaires autorisés sont les déplacements d'un carreau à droite ou d'un carreau vers le haut. Dessiner tous les chemins possibles de A à B(3,3). Comment pouvait-on prévoir leur nombre ?

Si B est le point de coordonnées (3,3), quelle est la probabilité pour que le chemin (A,B), passe par C(3,2) ? par D(1,2) ?

Cet exercice et tous ceux qui lui ressemblent ont pour but de mettre en relation la grille rectangulaire (n,p) et le nombre de chemins d'un coin au coin opposé qui vaut C_{n+p}^{p} ou encore évidemment C_{n+p}^{n}. Le nombre de chemins partant de A, atteignant les points du maillage s'établit selon la même règle que la construction du triangle de Pascal.

Utilisons maintenant cette vision bien connue dans un exercice moins banal, avec un appareillage théorique réduit.

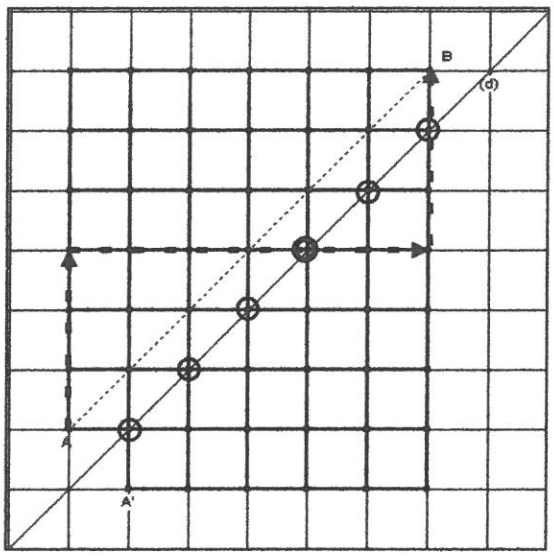
□ **Exercice 2**

L'entrée à un spectacle pour enfants coûte 11 francs. Certains enfants viennent avec une pièce de 10f et une pièce de 1f alors que les autres viennent avec une pièce de 10f et une pièce de 2f. Le caissier doit accueillir 100 enfants, et n'a pas prévu de monnaie en fond de caisse. Quelle est la probabilité pour que ce caissier délivre ses entrées sans problème, sachant qu'au bout du compte la caisse ne contient aucune pièce de 1f.

Quelques remarques :

La vente s'est bien passée si le caissier ne s'est pas trouvé en déficit de pièces de 1f pour rendre la monnaie. Pour 100 enfants, il aura reçu autant de pièces de 1f qu'il en aura rendu.

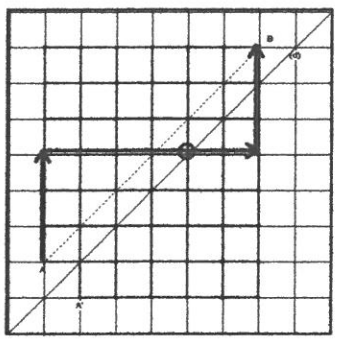
Le problème est le même en choisissant un nombre pair d'enfants, 2n.



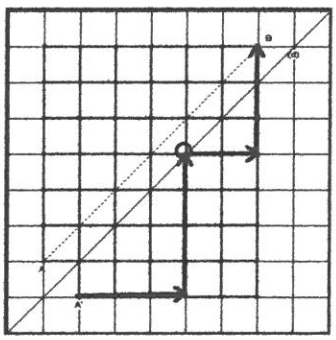
La connaissance préalable de l'image géométrique des combinaisons donne une clef pour la solution : sur un quadrillage (n,n) dans un carré allant de A(0,0) à B(n,n), prenons comme déplacements élémentaires, un carreau vers la droite pour une pièce de 1f rendue, et un carreau vers le haut pour une pièce de 1f reçue. Une caisse avec ou sans déficit en pièces de 1f est représentée par un chemin de A à B . Il y en a donc C_{2n}^{n}.

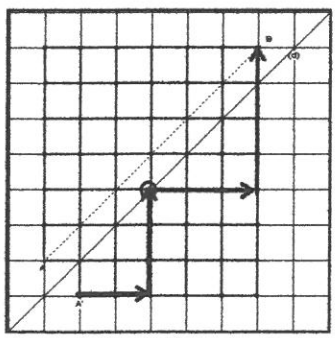
Les caisses déficitaires sont celles qui correspondent à des chemin franchissant la première diagonale, c'est à dire passant par l'un des points cerclés.

Dénombrons ces chemins en utilisant une symétrie partielle par rapport à la droite (d) des points cerclés de la manière suivante :

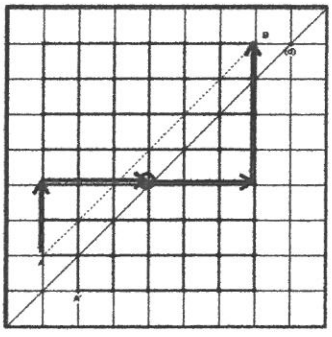


pour tout chemin rencontrant (d) en un premier point cerclé, on construit le chemin de A' symetrique de A par rapport à (d) à B, en symétrisant la partie du chemin initial comprise entre A et le premier point cerclé.





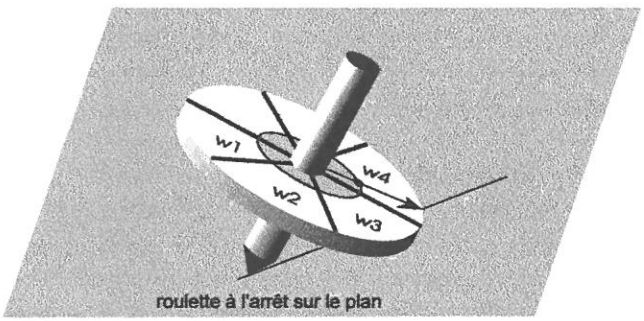
A tout chemin de A' à B (et passant donc obligatoirement par au moins un point cerclé), le procédé réciproque associe un chemin de A à B associé à une caisse déficitaire.

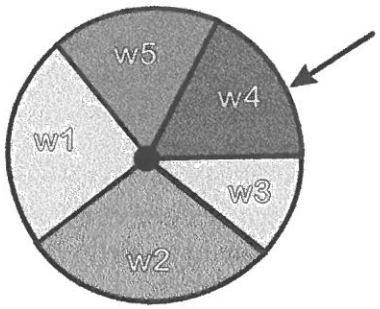


Le nombre de caisses déficitaires est donc le nombre de chemins de A' à B, c'est-à dire $C_{(n-1)+(n+1)}^{n-1}$, soit C_{2n}^{n-1}.

La probabilité cherchée est donc p= 1- C_{2n}^{n-1}/C_{2n}^{n} = 1/n+1.

2. *Utilisation du modèle des roulettes :*





la figure représente une roulette de circonférence unité, qui après rotation s'arrête de manière aléatoire sur le plan. La zone en contact avec le plan est alors w_1 ou w_2 ou w_3...et cet arrêt se produit avec une probabilité p_1 ou p_2 ou p_3... égale à la longueur de l'arc de circonférence de chaque zone.

Le fait d'actionner la roulette constitue une épreuve. L'ensemble de tous les événements élémentaires ou éventualités ω_n possibles est l'univers U.

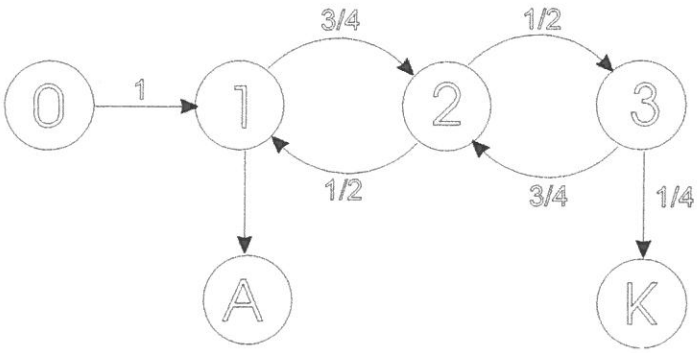
De manière évidente $\sum_{n\geq 1} p_n = 1$ et $p_n \geq 0$. On a aussi $U=\{\omega_1, \omega_2, \ldots, \omega_k\}$. On suppose $k \in N$. Le vecteur $p=\{p_1, p_2, \ldots, p_k\}$ est une probabilité sur U. Le couple (U, p) est un espace probabilisé discret fini. Si on considère l'expérience consistant à actionner un certain nombre de fois i la roulette, l'univers correspondant est U^i dont les éléments sont les n-mots de l'alphabet U.

Si l'on actionne i fois la roulette, l'éventualité ω_n se produit F_n fois. F_n est la fréquence absolue et $f_n = F_n/i$ est la fréquence relative de l'événement élémentaire ω_n. On supposera la roulette « bien équilibrée et symétrique », ce qui permet de conjecturer que pour i grand, $p_n \approx f_n$. Toutes les expériences confirment cette conjecture et la qualité de l'approximation de p_n par f_n s'améliore avec la croissance de i. p_n sera donc souvent « déterminée » expérimentalement par la valeur observée pour f_n.

3. *Promenades*

Des processus qui sont simulés d'une façon ou d'une autre par des roulettes sont des *processus aléatoires discrets*. Les étudier, c'est aborder le domaine des *probabilités discrètes*. Chacun de ces processus peut être schématisé par un arbre sur lequel un chemin décrit une des différentes façons de le réaliser. On s'intéresse évidemment aux probabilités des différentes exécutions possibles du processus.

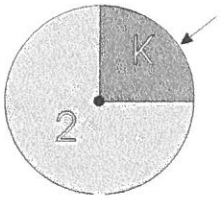
Un processus aléatoire discret peut être représenté par un graphe orienté tel celui de la figure :



Ses noeuds correspondent à des *états*.

A chaque état correspond une roulette qui définit la probabilité d'accéder à l'un des états suivant du graphe.

Si, par exemple, on se trouve à l'état 3, alors la roulette de la figure ci-contre gouverne l'accès aux états suivants possibles 2 ou K.



On appelle *promenade aléatoire* un tel cheminement, gouverné par des roulettes, sur un graphe orienté. On peut aller jusqu'à dire que le calcul des probabilités discrètes est l'étude de promenades aléatoires sur des graphes orientés.

Deux règles s'appliquent : la probabilité d'un chemin est égale au produit des probabilités des arcs parcourus, et la probabilité de réaliser un événement A est la somme des probabilités des chemins aboutissant à la réalisation de A.

Voici quelques exemples d'exercices proposés par A. Engel :

A propos de statistiques :

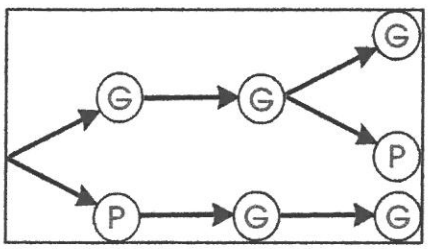
Le *recueil*, la *description*, le *traitement*, la *signification de données* constituent l'objectif de la *statistique*. Un problème central consiste à recueillir des données issues de processus gouvernés par des roulettes inconnues et d'en déduire une connaissance plus ou moins fiable de ces roulettes.

□ **Exercice 1**

Adam, Eve et leur fils Abel au tennis

Adam dit à son fils Abel : « Je te donnerai plus d'argent de poche si des trois parties que tu vas jouer, tu en gagnes deux de suite. Tu joueras alternativement contre moi et ta mère Eve ». Abel gagne généralement plus facilement contre Adam que contre Eve. Par quelle rencontre doit-il commencer ?

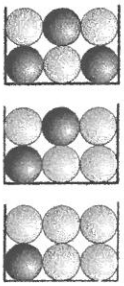
Si l'on note a la probabilité de victoire contre Adam et e la probabilité de victoire contre Eve, le graphe orienté du gain de deux parties successives est le suivant :



Si l'on pondère chaque arc par la probabilité associée, soit quand la première partie se joue contre Adam, soit quand elle se joue contre Eve, la probabilité de gain maximum pour Abel est dans le premier cas ae(2-a) et dans le second ae(2-e). Il est facile de conclure qu'il doit choisir d'affronter d'abord son père.

□ **Exercice 2**

Le voleur de Bagdad



Le calife de Bagdad faisait jeter les voleurs dans un cul-de-basse-fosse, mais leur laissait au préalable une dernière chance de rester libre. Chacun avait le droit de choisir une seule boule dans l'une des trois urnes de la figure ci-contre. Avec le tirage d'une boule blanche, il était libre.

Un jour un voleur plus malin, demande au calife s'il pouvait, avant le tirage, changer la répartition des boules dans les urnes. Le calife pensant que cela ne changerait rien à ses chances, accéda à sa demande...

Un graphe permet là aussi facilement de conclure.

□ **Exercice 3**

La météo en Syldavie

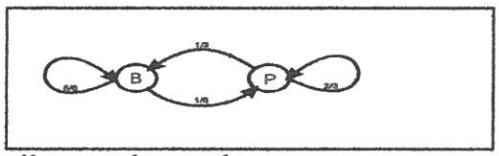
La figure 1.22 montre l'évolution du temps en Syldavie. S'il fait beau (B) alors il fera encore beau le lendemain avec une probabilité de 5/6. S'il pleut (P) alors il pleuvra le lendemain avec une probabilité de 2/3. On est dimanche, il fait beau. Quelle est la probabilité:

a) qu'il fasse beau le mardi;

b) qu'il pleuve mercredi;

c) qu'il fasse beau le jeudi ?

Une étude de cheminements sur le graphe orienté:



permet facilement de conclure

Statistiques et probabilités.

A. Engels propose parmi beaucoup d'autres, les trois exercices suivants, qui situent bien les problèmes que nous évoquerons par la suite.

□ **Exercice 1**

Le Président et le Journaliste. A vrai dire qui a raison?

En Syldavie tout le monde ment avec une probabilité **p** sauf deux personnes qui disent toujours la vérité: le Président et le présentateur du journal télévisé. Le Président a décidé de se présenter pour un nouveau mandat. Cette nouvelle parvient au présentateur par l'intermédiaire de **n** personnes successives. Quelle est la probabilité pour qu'il annonce une nouvelle exacte ?

□ **Exercice 2**

Le douanier. Doué ou chanceux ?

Les neuf passagers d'un avion se répartissent en cinq honnêtes personnes et quatre escrocs. Un douanier fouille trois passagers. Il trouve trois escrocs. Que dire des deux hypothèses :

1) le douanier a eu de la chance, « l'échantillon est aléatoire »

2) le douanier est doué, son flair ou d'autres éléments non liés au hasard l'ont guidé vers les escrocs.

Réexaminer ces hypothèses quand en fouillant quatre personnes il ne trouve que trois escrocs.

□ **Exercice 3**

Les taxis

A Paris, les taxis sont numérotés 1, 2, 3, ..., T, mais nous ne connaissons pas T.

Un chauffeur prétend que $T \geq 3000$. Dans la rue vous observez quatre taxis portant les numéros 512, 987, 355 et 1200.

1) Hypothèse A : $T \geq 3000$

Hypothèse B : $T < 3000$

Que pensez-vous de ces deux hypothèses ?

2) Peut-on estimer le nombre de taxis numérotés à Paris:

- ponctuellement ?

- par un intervalle ?

CORRELATION INDEPENDANCE

Variables aléatoires liées

Nous avons tous rencontré des assertions du genre : plus une personne a un niveau d'études élevé, plus elle a de chances, devenue adulte, d'avoir un emploi. Plus l'année a été pluvieuse et plus la circonférence de l'anneau de croissance d'un tronc d'arbre a de chance d'être grande. Plus il y a de chômage, plus il y a de délinquance juvénile. Chacune de ces assertions concerne deux variables aléatoires et affirme qu'il y a une relation entre elles. Ces assertions indiquent aussi que cette relation est d'un type spécial, appelé *corrélation*. Dans ce paragraphe on se propose de formuler avec précision le concept de corrélation, de façon à mieux comprendre la signification de telles assertions.

Variables aléatoires indépendantes

Considérons le lancer d'une pièce régulière trois fois de suite. Chaque événement élémentaire a une probabilité de 1/8. On considère les variables aléatoires suivantes :

X est le gain d'un joueur qui reçoit 3f s'il amène face au premier lancer, plus 2f s'il amène face au deuxième lancer, plus 1f s'il amène face au troisième lancer. Y est le nombre de faces obtenues, U est l'indicatrice de face au premier lancer (0 si pile, 1 si face) et V l'indicatrice de face au second lancer.

lancers	X	Y	U	V
FFF	6	3	1	1
FFP	5	2	1	1
FPF	4	2	1	0
PFF	3	2	0	1
FPP	3	1	1	0
PFP	2	1	0	1
PPF	1	1	0	0
PPP	0	0	0	0

Dressons les tableaux de probabilités conjointes des couples de variables (X, Y) et (U, V).

X \ Y	0	1	2	3	p(X=x)
0	1/8	0	0	0	1/8
1	0	1/8	0	0	1/8
2	0	1/8	0	0	1/8
3	0	1/8	1/8	0	2/8
4	0	0	1/8	0	1/8
5	0	0	1/8	0	1/8
6	0	0	0	1/8	1/8
p(Y=y)	1/8	3/8	3/8	1/8	1

U \ V	0	1	p(U=u)
0	1/4	1/4	1/2
1	1/4	1/4	1/2
p(V=v)	1/2	1/2	1

Deux événements **E** et **F** sont indépendants si en termes de probabilités, **$P(E\cap F)=P(E).P(F)$.** Appliquons-le aux événements énumérés dans le tableau de probabilités conjointes du couple (X,Y). Choisissons comme événement E l'événement X=3. Sa probabilité est 2/8. L'événement F sera l'événement Y=1. Sa probabilité est de 3/8. Ces deux événements sont indépendants si et seulement si la probabilité de leur intersection satisfait la loi de multiplication, c'est-à-dire si P(X=3,Y=1) = (2/8)x(3/8) soit 3/32. Cependant, P(X=3,Y=1) est indiqué dans la case correspondante et vaut 1/8. Donc elle ne suit pas la règle de multiplication et les deux événements X=3 et Y=1 ne sont pas indépendants. De même le tableau indique que quelles que soient les valeurs x de X et y de Y indiquées dans le tableau, les événements X=x et Y=y ne sont pas indépendants.

Nous trouvons une situation différente dans le tableau de probabilité conjointe au couple (U,V). Là on trouve que P(U=0)=1/2, P(V=0)=1/2, et P(U=0,V=0)=1/4, donc que la règle de multiplication est satisfaite. Par conséquent les événements U=0 et V=0 sont indépendants. De même, U=0 et V=1 sont indépendants, U=1 et V=1 aussi. Dans une telle situation on dit que les deux variables U et V sont indépendantes, en accord avec la définition suivante : *deux variables aléatoires X et Y sont indépendantes, si, quelles que soient les valeurs x de X et y de Y, les événements X=x et Y=y sont indépendants.* La règle de multiplication pour les événements indépendants nous donne tout simplement une méthode pour reconnaître des variables aléatoires indépendantes. Deux variables aléatoires X et Y sont indépendantes si et seulement si chaque nombre des cases du tableau de probabilité du couple (X,Y) est le produit des probabilités marginales correspondantes. Les variables aléatoires qui ne sont pas indépendantes sont dites *dépendantes.*

La définition de variables aléatoires indépendantes s'étend facilement à trois ou plusieurs variables. Elles sont caractérisées par la règle de multiplication :

$$P(A=a,B=b,...,Z=z) = P(A=a).P(B=b)...P(Z=z).$$

Covariance de deux variables aléatoires

Alors que l'espérance mathématique de la somme de deux variables aléatoires est égale à la somme de leurs espérances mathématiques, l'espérance mathématique de leur produit n'est pas forcément le produit de leurs espérances mathématiques. Puisque E(XY) et E(X).E(Y) peuvent différer,

intéressons-nous à leur différence et voyons quel renseignement elle nous donne sur la liaison entre X et Y. On appelle covariance de X et de Y le nombre noté Cov(X,Y) = E(XY) - E(X).E(Y).

En notant $E(X) = \mu_X$ et $E(Y) = \mu_Y$, il vient $Cov(X,Y) = E(XY) - \mu_X \mu_Y$.

Calculons $E[(X-\mu_X)(Y-\mu_Y)] = E(XY-\mu_X Y-\mu_Y X+\mu_X \mu_Y) = E(XY-\mu_X Y-\mu_Y X)+\mu_X \mu_Y$

En définitive, on obtient $E[(X-\mu_X)(Y-\mu_Y)] = E(XY) - \mu_X \mu_Y$ ce qui est égal à Cov(X,Y). Par conséquent : $Cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]$.

Evaluons maintenant $Cov(X-\mu_X, Y-\mu_Y) = E[(X-\mu_X)(Y-\mu_Y)] - (\mu_{X-\mu_X})(\mu_{Y-\mu_Y})$. Or $\mu_{X-\mu_X} = \mu_{Y-\mu_Y} = 0$

Donc $Cov(X-\mu_X, Y-\mu_Y) = E[(X-\mu_X)(Y-\mu_Y)] = Cov(X,Y)$.

Une mesure de corrélation.

Considérons l'égalité $Cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]$. En examinant cette espérance mathématique, nous aurons des indications sur le comportement de X et Y l'un par rapport à l'autre. Pour obtenir ces indications, on remarque d'abord que si une variable aléatoire n'est pas constante et si toutes ses valeurs non nulles sont positives, alors sa valeur moyenne est positive. De même, si une variable aléatoire n'est pas constante et que toutes ses valeurs non nulles sont négatives alors sa valeur moyenne est négative.

Pour observer le comportement des variables X et Y, on considère les couples de valeurs (x,y) associées au même élément de l'univers des possibles. Nous les appellerons *valeurs correspondantes*. Supposons que X et Y aient la propriété que, si elles diffèrent de leurs valeurs moyennes respectives, les valeurs correspondantes de X et Y sont, soit toutes deux supérieures, soit toutes deux inférieures à leurs valeurs moyennes. Donc $X-\mu_X$ et $Y-\mu_Y$, quand elles ne sont pas nulles sont ou toutes deux positives ou toutes deux négatives pour des couples correspondants de valeurs (x,y). Par conséquent, leur produit, s'il n'est pas nul est positif et l'espérance mathématique de leur produit, qui est égal à Cov(X,Y) est aussi positif. Supposons, d'autre part, que X et Y aient la propriété que les valeurs de X supérieures à la moyenne de X correspondent aux valeurs de Y inférieures à la moyenne de Y et inversement. Alors $X-\mu_X$ et $Y-\mu_Y$, quand elles ne sont pas nulles ont des signes opposés. Donc leur produit s'il n'est pas nul, est toujours négatif et l'espérance mathématique de leur produit qui est égal à Cov(X,Y) est aussi négatif. En bref, lorsque les valeurs correspondantes de X et Y s'écartent dans le même sens de leur valeur moyenne Cov(X,Y) est positif et quand elles s'en écartent en sens opposé, Cov(X,Y) est négatif. Ceci nous incite à utiliser Cov(X,Y) comme mesure de la façon dont sont reliés l'écart de X et celui de Y. De façon à rendre cette mesure indépendante des unités avec lesquelles sont mesurées X et Y, on utilisera plutôt Cov(X*,Y*) où X* et Y* sont les variables aléatoires centrées réduites de X et Y, à condition que $\sigma_X \neq 0$ et $\sigma_Y \neq 0$.

Le *coefficient de corrélation* de X et Y se note r(X,Y) et se définit par:

si $\sigma_X = 0$ ou $\sigma_Y = 0$, $r(X,Y) = 0$

si $\sigma_X \neq 0$ et $\sigma_Y \neq 0$, $r(X,Y) = Cov(X^*,Y^*)$

Si $r(X,Y) \neq 0$, on dit que X et Y sont *corrélées*. Si $r(X,Y) = 0$ elles ne le sont pas. Pour les calculs il sera pratique d'avoir l'expression de r(X,Y) en fonction de X et Y plutôt que de X* et Y*. Si $\sigma_X \neq 0$ et $\sigma_Y \neq 0$ on a :

$$r(X,Y) = Cov(X^*,Y^*) = E(X^*.Y^*) - \mu_{X^*}\mu_{Y^*} = E(X^*.Y^*)$$

$$r(X,Y) = E(X^*.Y^*) = E\left(\frac{(X-\mu_X)}{\sigma_X}\frac{(Y-\mu_Y)}{\sigma_Y}\right) = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X\sigma_Y} = \frac{Cov(X,Y)}{\sigma_X\sigma_Y}.$$

Différentes valeurs du coefficient de corrélation

De façon à déterminer l'ordre de grandeur des valeurs possibles de r(X,Y), on va déterminer une formule pour Var(X+Y).

D'après la définition de la variance, Var(X+Y) est la moyenne du carré de l'écart de X+Y. Donc on aura le résultat cherché en trois étapes : d'abord on détermine l'écart de X+Y, on élève au carré, puis on cherche la valeur moyenne du carré de l'écart. L'écart de X+Y est :

$X+Y - E(X+Y) = X+Y- E(X) -E(Y) = X+Y - \mu_X - \mu_Y = (X - \mu_X) + (Y - \mu_Y)$.

En élevant au carré on obtient : $(X - \mu_X)^2 + (Y - \mu_Y)^2 + 2(X - \mu_X)(Y - \mu_Y)$.

La valeur moyenne de cette expression est :

$$E[(X - \mu_X)^2 + (Y- \mu_Y)^2 + 2(X- \mu_X)(Y- \mu_Y)] = E[(X- \mu_X)^2]+E[(Y- \mu_Y)^2]+ 2E[(X- \mu_X)(Y- \mu_Y)].$$

Mais $E[(X- \mu_X)^2] = Var(X)$, $E[(Y- \mu_Y)^2] = Var(Y)$ et $E[(X- \mu_X)(Y- \mu_Y)] = Cov(X,Y)$.

On a donc la formule : $Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$.

En appliquant cette formule aux variables régularisées X* et Y*, on obtient :

$$Var(X^*+Y^*) = Var(X^*) +Var(Y^*) + 2Cov(X^*,Y^*).$$

Mais $Var(X^*) = Var(Y^*) =1$ et $Cov(X^*,Y^*) = 2r(X,Y)$. Par conséquent $Var(X^*+Y^*)=2(1+ r(X,Y))$. D'après la définition de la variance nous savons qu'elle est toujours positive et donc $r(X,Y) \geq -1$. En établissant que $Var(X-Y) = 2(1- r(X,Y))$, on montre que $r(X,Y) \leq 1$.

En définitive, on a obtenu $-1 \leq r(X,Y) \leq 1$.

Valeurs extrêmes du coefficient de corrélation

La valeur absolue du coefficient de corrélation de deux variables aléatoires indique quel est leur degré de corrélation. Sa valeur maximum est 1 et est obtenue lorsque r(X,Y) atteint l'une de ses bornes, -1 ou 1. Pour avoir une image intuitive de la signification d'un coefficient de corrélation, examinons à quoi correspondent ces valeurs extrêmes.

Comment X et Y sont-ils reliés l'un à l'autre si r(X,Y)=1 ?

Pour répondre à cette question, rappelons-nous d'abord que chacune des variables aléatoires X et Y affecte une valeur définie de x et y, à chaque élément de l'univers des possibles sur lequel les variables aléatoires sont définies. Il peut arriver que des événements élémentaires de l'univers des possibles aient une probabilité nulle. Nous ne les considérerons pas parce qu'une variable obtenue avec la probabilité 0 n'est pas une valeur possible et par conséquent ne nous concerne pas.

Nous avons vu dans un paragraphe précédent que $Var(X^*-Y^*) = 2(1-r(X,Y))$. Si r(X,Y)=1, il s'en suit que $Var(X^*-Y^*) = 0$. Si la variance d'une variable aléatoire est 0, cela signifie que l'écart de chaque valeur possible est nul. C'est-à-dire qu'aucune valeur possible de la variable ne diffère de sa moyenne. Par conséquent pour des valeurs possibles, $X^*-Y^* = E(X^*-Y^*)$. Mais $E(X^*-Y^*) = E(X^*) -E(Y^*) = 0 - 0 = 0$. Donc, par conséquent pour toutes les valeurs possibles de X et de Y, $X^*-Y^*=0$, ou $X^* = Y^*$. En se souvenant de la définition d'une variable aléatoire réduite, on voit que :

$$\frac{X - \mu_X}{\sigma_X} = \frac{Y - \mu_Y}{\sigma_Y}$$

et la résolution de cette équation en Y fonction de X donne : $Y = \frac{\sigma_X}{\sigma_Y} X + \frac{\sigma_X \mu_Y - \sigma_Y \mu_X}{\sigma_X}$.

Cette équation est de la forme Y = mX + p. De plus puisque r(X,Y) = 1, on a $\sigma_X \neq 0$ et $\sigma_Y \neq 0$, et donc tous deux positifs, et m aussi : la droite « monte ». Cela signifie que si X augmente, Y aussi. Un raisonnement analogue montre que si r(X,Y) = -1 on obtient une droite qui « descend ».

On a montré que si r(X,Y) = ±1, alors pour les valeurs possibles de X et Y, Y = mX + b ou m ≠ 0. Démontrons maintenant la réciproque. Supposons que Y = mX + b avec m ≠ 0. $E[(X-\mu_X)(Y-\mu_Y)] = Cov(X,Y)$. Puisque Y = mX + b on a $\mu_Y = m\mu_X + b$. Par conséquent $Y - \mu_Y = m(X - \mu_X)$. En effectuant cette substitution on obtient $Cov(X,Y) = E[m(X-\mu_X)^2] = mE[(X-\mu_X)^2] = m\sigma_X^2$.

D'autre part $\sigma_Y = |m|\sigma_X$. Par conséquent : $r(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \frac{m\sigma_X^2}{\sigma_X |m| \sigma_X} = \pm 1$ suivant que m est positif ou négatif.

Erreurs à éviter sur la corrélation

On fait souvent deux erreurs sur la signification du mot corrélation. La première fait du mot *non corrélées*, un synonyme de *non liées*. Si X et Y sont des variables aléatoires, et $Y = X^2$, X et Y sont très certainement liées bien que non corrélées.

L'autre erreur très répandue est de confondre *corrélation* et *causalité*. Lorsque r(X,Y) a une valeur voisine de 1 ou de -1, permettant de dire que la corrélation entre X et Y est forte, cela signifie que les fluctuations de X et de Y sont plus ou moins simultanées.

Si r est positif, cela signifie qu'un accroissement de la valeur de X est en général accompagné d'un accroissement de celle de Y. Si r est négatif, l'accroissement de X est accompagné en général d'une diminution de celle de Y. Mais cela ne signifie pas que la variation de X est nécessairement la cause de la variation de Y. En fait, il existe plusieurs situations différentes compatibles avec une forte corrélation de X et Y. Par exemple,

1- la variation de X peut causer la variation de Y;

2- la variation de Y peut causer celle de X;

3- les variations de X et de Y sont toutes deux les effets d'une autre cause opératoire;

4- la liaison apparente des variations n'est qu'une coïncidence.

A cause de toutes ces possibilités, une corrélation doit toujours être interprétée avec prudence. Une forte corrélation incite à considérer qu'il puisse y avoir une relation causale, mais elle ne prouve pas qu'une telle relation existe.

La phrase citée au départ, nous donne un exemple de corrélation mal interprétée. Il existe une corrélation positive entre le nombre d'années qu'un adulte a passées à l'école et son revenu annuel. On cite souvent ce fait pour prouver l'assertion par laquelle l'instruction accroît la capacité de gagner de l'argent pour un individu. Cependant ce fait isolé ne prouve pas absolument l'assertion. Il est possible, par exemple, qu'une forte capacité de gain et qu'un haut niveau d'instruction soient tous deux les conséquences du fait d'être né dans une famille au statut économique et social élevé. Il peut être vrai que l'instruction permet de gagner davantage d'argent, mais il faut des critères supplémentaires pour l'affirmer.

□ **Exercice**

On range trois boules dans trois tiroirs T1, T2, T3, chaque tiroir pouvant contenir jusqu'à trois boules. X1 représente le nombre de boules dans T1 et N le nombre de tiroirs occupés.

1) Loi conjointe du couple (N,X1)

On a $N(\Omega) = \{1,2,3\}$, $X1(\Omega) = \{0,1,2,3\}$

N \ X1	0	1	2	3	Loi de N
1	2/27	0	0	1/27	1/9
2	6/27	6/27	6/27	0	2/3
3	0	6/27	0	0	2/9
Loi de X1	8/27	4/9	6/27	1/27	1

Donnons quelques explications :

a) Sur la présence de 0 dans le tableau :

Par exemple $P((X1=1)\cap(N=1)) = 0$: s'il n'y a qu'une boule dans le tiroir T1, il faut nécessairement au moins deux tiroirs occupés. De même $P((X1=3)\cap(N=2)) = 0$, car si trois boules sont dans T1, il est le seul occupé. Nous laissons au lecteur le soin de vérifier les autres 0!

b) Sur le calcul des probabilités :

Le nombre de dispositions possibles est le nombre d'applications d'un ensemble à trois éléments dans un ensemble à trois éléments, c'est-à-dire 27 dispositions équiprobables. (On numérote les boules 1, 2, 3).

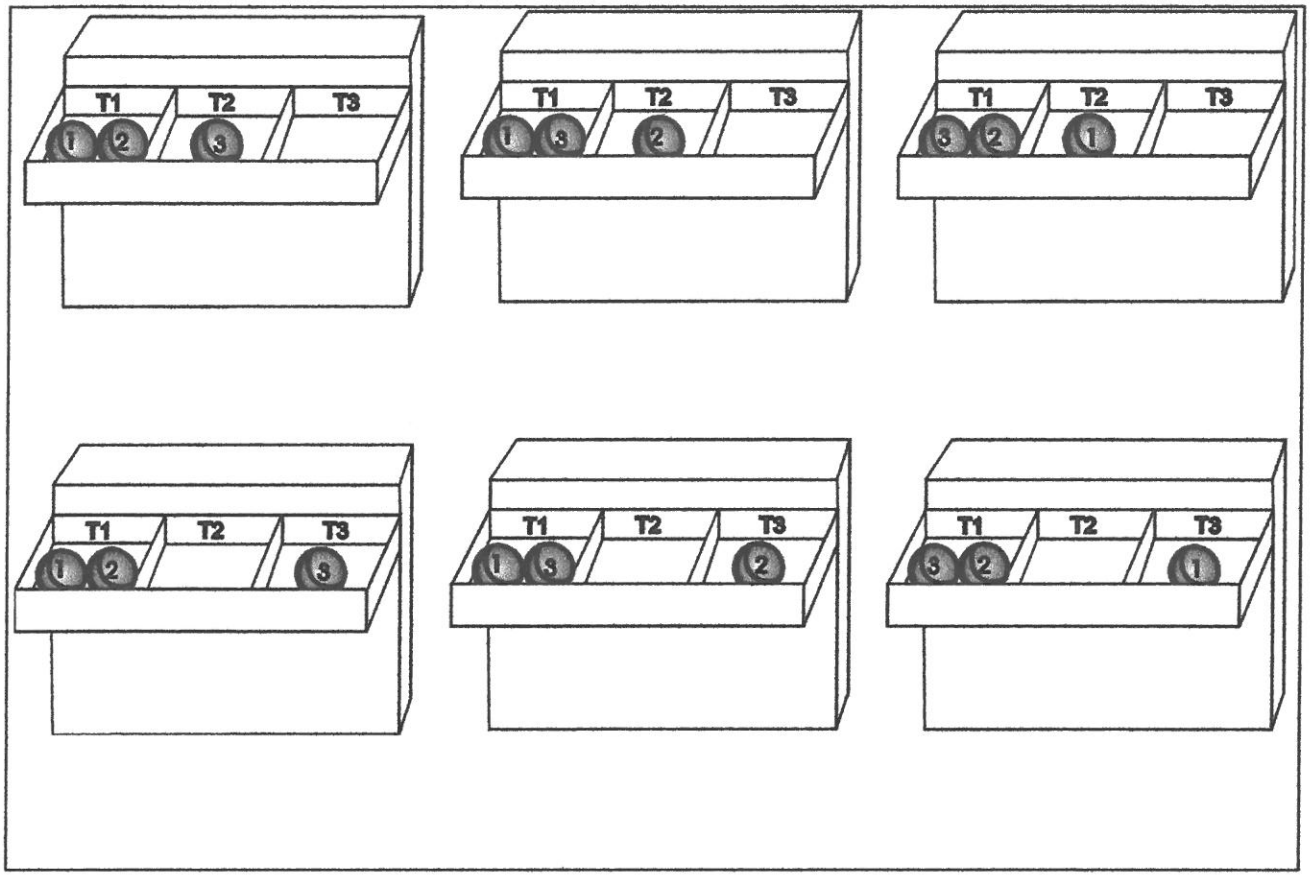
Par exemple $P((X1=0)\cap(N=1)) = 2/27$, car il n'y a qu'un tiroir occupé et T1 ne contient aucune boule. Ainsi les trois boules sont dans T2 ou les trois boules sont dans T3, deux cas favorables.

De même $P((X1=2)\cap(N=2)) = 6/27$, car il y a deux boules dans T1 et deux tiroirs occupés ce qui donne 6 configurations possibles comme indiqué sur le dessin ci-après.

2) Les lois marginales de X1 et N sont indiquées dans le tableau, ce qui donne :

$$E(X1) = \frac{12}{27} + 2\times\frac{6}{27} + 3\times\frac{1}{27} = 1 \qquad \text{et} \qquad E(N) = \frac{1}{9} + 2\times\frac{2}{3} + 3\times\frac{2}{9} = \frac{19}{9}$$

on a $E(X1) = 3\times\frac{1}{3} = 1$ car X1 suit la loi binomiale de paramètres 3 et 1/3.



3) L'espérance de X1.N est donnée par la formule :

$$E(X1.N) = \sum_{i=1}^{3} \sum_{j=1}^{3} i.j.p_{ij} = \frac{12}{27} + \frac{18}{27} + \frac{24}{27} + \frac{3}{27} = \frac{19}{9}$$

4) On peut alors faire deux constatations :

$Cov(X1,N) = E(X1.N) - E(X1).E(N) = 0$

$P((N=1) \cap (X1=1)) = 0 \neq P(N=1).P(X1=1)$

N et X1 ne sont donc pas indépendantes, bien que leur corrélation soit nulle.

ECHANTILLONNAGE

Echantillon

Quand la population comporte un grand nombre d'individus, on ne peut pas ou on ne veut pas, en général pour des raisons économiques, les examiner tous. Les observations ne porteront donc que sur un nombre restreint d'individus qu'il faudra choisir. Les individus sélectionnés constitueront un *échantillon*, leur nombre sera la *taille* de l'échantillon.

Le problème que nous nous posons est de tirer de l'information fragmentaire résultant de l'inspection d'un échantillon des informations valables pour l'ensemble de la population que nous désirons étudier. Cette population est dite : « *population mère* ».

Ce n'est que dans des cas très particuliers que l'on pourra être sûr des caractéristiques de la population mère. En général, il faudra s'exprimer en termes de probabilités.

Nous allons commencer par étudier un échantillon, puis nous examinerons dans quelles conditions on peut en déduire des estimations valables pour la population mère.

Comment prélever l'échantillon ?

Nous avons déjà vu qu'il était essentiel que l'échantillon soit prélevé au hasard. Cela signifie que chaque individu doit avoir une chance calculable d'être sélectionné. L'adjectif « calculable » suppose deux points :

1. Cette chance n'a pas besoin d'être calculée. Il suffit qu'elle soit calculable, c'est-à-dire que le processus choisi simule une loi de probabilité.

2. Cela implique que chaque individu ait une chance égale. C'est le cas si chaque tirage est indépendant ; on dit alors que l'on a affaire à un échantillonnage simple. Mais parfois la recherche de l'efficacité dans la préparation des expériences conduit à abandonner ce type d'échantillon sans que cela empêche d'utiliser les règles du calcul des probabilités.

Finalement on peut se demander comment procéder pour que l'échantillon soit prélevé au hasard.

Cela est extrêmement difficile. Emile Borel a même démontré que l'homme était incapable de reproduire le hasard. En aucun cas on ne doit se fier à son intuition, ni à son flair pour essayer d'imiter le hasard. Un grand nombre d'échecs s'expliquent parce que certains expérimentateurs transgressent cette règle.

Le moyen le moins imparfait consiste à utiliser une table de nombre au hasard. Pour cela, on affecte un numéro matricule à chacun des individus qui constituent la population, puis on sélectionne un échantillon en utilisant une telle table.

Ce mode opératoire qui est toujours correct, peut subir quelques modifications dans certains cas.

Si l'effectif de la population est extrêmement important, avant d'échantillonner en appliquant brutalement la méthode précédente, il faut réfléchir aux conséquences matérielles.

Parfois, bien que la population soit pratiquement infinie, il n'y a aucune difficulté : supposons qu'un métallurgiste veuille mesurer l'épaisseur des tôles qui sortent d'un laminoir. Il pourra mesure l'épaisseur en certains points dont les coordonnées auront été tirées au hasard en utilisant une table.

Il en va tout autrement si les individus sont géographiquement très dispersés : par exemple, l'ensemble de la population française. Pour faire une enquête, on pourrait tirer au hasard des individus en utilisant, par exemple, leur numéro de sécurité sociale. Ce procédé, théoriquement parfait, nécessiterait certainement d'envoyer des enquêteurs dans tous les départements français. En fait, on opère autrement : on tire au sort quelques départements, puis dans chaque département des arrondissements, etc... et enfin des individus.

Ces procédés d'échantillonnage ont donné lieu au développement d'une théorie très importante, celle des *sondages*, qui a permis d'élaborer des schémas plus ou moins complexes qui fournissent pour une dépense donnée, le maximum d'informations.

Echantillon bernoulliens et exhaustifs

Les échantillons sont tirés au hasard dans la population mère. Si la population est finie et comporte un nombre N d'individus, deux types de tirages peuvent être envisagés.

Les tirages bernoulliens ou non exhaustifs supposent que tout individu tiré est replacé, après observation du caractère, dans la population mère. Il peut donc être tiré plusieurs fois.

Les tirages exhaustifs ou sans remise supposent que tout individu tiré ne peut jamais être retiré une deuxième fois ; ceci revient à tirer n individus simultanément, n « N.

Indépendance des observations

Les observateurs se demandent souvent si les observations successives qu'ils font - d'un même phénomène - sont indépendantes ou non. C'est en fait une question extrêmement importante. Nous allons essayer d'y apporter quelque clarté.

Il faut d'abord prendre conscience que l'indépendance statistique ne résulte pas de considérations théoriques ; ce n'est pas quelque chose qui se démontre à la manière d'un théorème de mathématique, c'est quelque chose qui résulte des observations. Ce sont donc les expérimentateurs et eux seuls qui peuvent décider si les observations successives sont indépendantes ou non, sous réserve que leurs arguments ne soient pas en contradiction avec la théorie.

LOIS LIMITES

Loi faible des grands nombres

Soit (X1, X2, ..., Xn) une suite de n variables aléatoires indépendantes de même espérance E(X) et même variance σ^2.

On pose Sn = X1 + X2 + ... + Xn et $\overline{X}n = \frac{1}{n}Sn$. Alors pour tout $\varepsilon > 0$, $P(|\overline{X}n - E(X)| < \varepsilon)$ tend vers 1 quand n tend vers l'infini.

Ce théorème justifie le fait d'attribuer, a priori, comme probabilité p = E(X) à un événement, sa fréquence statistique $\overline{X}n$ quand n est grand.

Théorème de la limite centrée

Si (X1, X2, ..., Xn) est une suite de variables aléatoires indépendantes de même loi de probabilité, d'espérance m et de variance σ^2; alors pour n grand, la loi de la moyenne $\overline{X}n = \frac{1}{n}Sn$ peut être approchée par une loi normale de moyenne m et d'écart type $\frac{\sigma}{\sqrt{n}}$. ($T = \frac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}}$ suit une loi normale centrée réduite). Dans ce théorème il n'est pas nécessaire de connaître la loi des Xi.

Approximation d'une loi binomiale par une loi normale

Si X suit une loi binomiale de paramètres n et p alors la loi de X est approchée par une loi normale de paramètres : espérance = np et variance =np(1-p).

Lois d'échantillonnage (non exhaustif)

- **La moyenne d'échantillon**

$\overline{X}n = \frac{1}{n}Sn$ suit une loi normale $N(m, \frac{\sigma}{\sqrt{n}})$ pour n grand.

m = moyenne de la population

σ = écart type de cette population.

- **La fréquence**

La population mère possède un pourcentage p d'éléments qui ont une propriété donnée. La variable aléatoire qui à tout échantillon de taille n associe la fréquence de la propriété observée, suit pour n grand, une loi normale $N(p, \sqrt{\frac{pq}{n}})$ où q = 1- p.

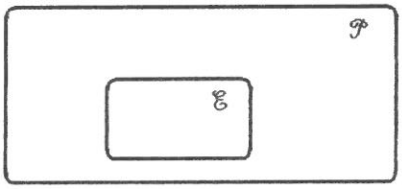
STATISTIQUES INFERENTIELLES

* * *

Estimation ponctuelle et par intervalle de confiance

Le problème

On veut étudier un paramètre, par exemple la moyenne ou l'écart type, d'un caractère quantitatif d'une population $\mathcal{P}$ (appelée population mère).



On travaille sur un échantillon $\mathcal{E}$ de cette population et on se pose la question :

Que peut-on déduire de la valeur observée du caractère dans $\mathcal{E}$ pour la valeur du paramètre dans $\mathcal{P}$? Avec quelle confiance ?

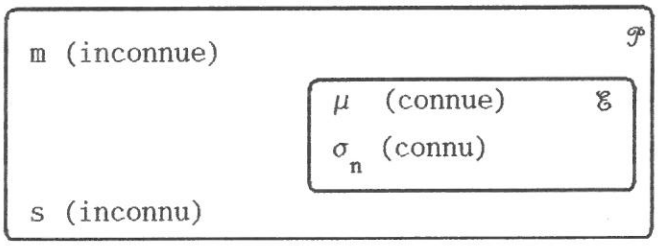
Hypothèses de travail

On suppose que l'effectif de la population est suffisamment grand pour que le prélèvement aléatoire d'un échantillon ne modifie pas notablement le caractère étudié et, ainsi, puisse être assimilé à un tirage avec remise.

A - Travail sur moyenne et écart type

I - Estimation ponctuelle

Problème



$\mathcal{E}$ est un échantillon aléatoire de taille n de $\mathcal{P}$. Le caractère a pour moyenne m et pour écart type s dans $\mathcal{P}$, pour moyenne μ et pour écart type σ_n dans $\mathcal{E}$. Quel lien y a-t-il entre m et μ ? entre s et σ_n ?

Définition

μ et $\sigma = \sqrt{\frac{n}{n-1}}\ \sigma_n$ sont des <u>estimations ponctuelles</u> respectives de m et de s.

Exercice 1 :

Un contrôle portant sur une machine emballant automatiquement et en série des paquets de beurre fournit les résultats suivants :

masse en g	247	248	249	250	251	252	253	254
nombre de paquets	2	6	8	13	11	5	3	2

Vérifier que 250,24 est une estimation ponctuelle de la moyenne m de la masse des paquets de beurre emballés et 1,67 une estimation ponctuelle de l'écart type s.

Exercice 2 : (Extrait B.T.S. Analyse Biologique 1984)

On tire au hasard, au sein d'une population $\mathcal{P}$ (très grande), un échantillon $\mathcal{E}$ de 100 sujets et l'on mesure la glycémie de chacun d'entre eux ; on obtient les résultats suivants :

glycémie dans l'intervalle (mg/100 ml)	effectif
[75 ; 80[	5
[80 ; 85[	10
[85 ; 90[	20
[90 ; 95[	36
[95 ; 100[	15
[100 ; 105[	8
[105 ; 110[	6

Estimer ponctuellement la moyenne et l'écart type de la glycémie dans la population $\mathcal{P}$. (On trouvera : $\mu = 92,20$ et $\sigma = 7,21$.)

Pourquoi cette estimation ponctuelle est-elle une bonne estimation ?

Appelons x_1, x_2, ..., x_n les valeurs observées pour un échantillon de taille n. En répétant les échantillons de taille n fixée, on définit des variables aléatoires X_1, X_2, ..., X_n.

Avec l'hypothèse de travail, les variables aléatoires X_1, X_2, ..., X_n sont indépendantes et de même loi de moyenne m et de variance s^2.

On a, pour tout i : $E(X_i) = m$ et $Var(X_i) = s^2$.

On va s'intéresser à :

* $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ — la moyenne d'échantillonnage,

* $V = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2$ — la variance d'échantillonnage.

$\bar{X}$ et V sont des variables aléatoires définies sur l'ensemble des échantillons de taille n de la population : ce sont des estimateurs.

Espérance et variance de $\bar{X}$

* $E(\bar{X}) = E\left(\frac{1}{n} \sum_{i=1}^{n} X_i \right) = \frac{1}{n} \sum_{i=1}^{n} E(X_i) = \frac{1}{n}\, n\, m = m.$

* $Var(\bar{X}) = V\left(\frac{1}{n} \sum_{i=1}^{n} X_i \right) = \frac{1}{n^2} \sum_{i=1}^{n} Var(X_i)$

car les X_i sont indépendantes.

Ainsi $Var(\bar{X}) = \frac{1}{n^2}\, n\, s^2 = \frac{s^2}{n}.$

Comme $E(\bar{X}) = m$, on dit que $\bar{X}$ est un estimateur sans biais de m ; comme de plus $\lim_{n \to +\infty} Var(\bar{X}) = 0$, on dit que $\bar{X}$ est un estimateur convergent de m.

(Ces notions ne figurent pas aux programmes des classes de S.T.S.)

Espérance et variance de V

* $E(V) = E\left[\frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2 \right] = \frac{1}{n}\, E\left[\sum_{i=1}^{n} (X_i - m + m - \bar{X})^2 \right]$

$$= \frac{1}{n} \sum_{i=1}^{n} E[(X_i - m)^2] + \frac{2}{n} \sum_{i=1}^{n} E[(X_i - m)(m - \bar{X})] + \frac{1}{n} \sum_{i=1}^{n} E[(m - \bar{X})^2]$$

$$= \frac{1}{n} \sum_{i=1}^{n} E[(X_i - E(X_i))^2] + \frac{2}{n}\, E\left[\sum_{i=1}^{n} [(X_i - m)(m - \bar{X})] \right]$$

$$+ \frac{1}{n} \sum_{i=1}^{n} E[(\bar{X} - E(\bar{X}))^2]$$

$$= \frac{1}{n} \sum_{i=1}^{n} Var(X_i) + \frac{2}{n}\, E\left[(m - \bar{X}) \sum_{i=1}^{n} (X_i - m) \right] + \frac{1}{n} \sum_{i=1}^{n} Var(\bar{X})$$

$$E(V) = \frac{1}{n} \sum_{i=1}^{n} s^2 + \frac{2}{n} E\left[(m-\bar{X})\ (n\bar{X} - nm)\right] + \frac{1}{n} \sum_{i=1}^{n} \frac{s^2}{n}$$

par définition de $\bar{X}$

$$= \frac{1}{n}\ n\ s^2 - 2\ E[(m-\bar{X})^2] + \frac{1}{n}\ n\ \frac{s^2}{n}$$

$$= s^2 - 2\ \mathrm{Var}(\bar{X}) + \frac{s^2}{n} = s^2 - 2\ \frac{s^2}{n} + \frac{s^2}{n} = \frac{n-1}{n}\ s^2.$$

V est donc un estimateur biaisé de s^2 ; par contre $E\left(\frac{n}{n-1}\ V\right) = s^2$, $\frac{n}{n-1}\ V$ est un estimateur sans biais de s^2.

* Dans le cas particulier où la distribution du caractère étudié dans $\mathcal{P}$ est normale, on montre que $\mathrm{Var}\left(\frac{n}{n-1}\ V\right) = \frac{2\ s^4}{n-1}$, ainsi $\lim_{n \to +\infty} \mathrm{Var}\left(\frac{n}{n-1}\ V\right) = 0$ et $\frac{n}{n-1}\ V$ est un estimateur convergent de s^2.

(Ce résultat ne se généralise pas à toutes les lois.)

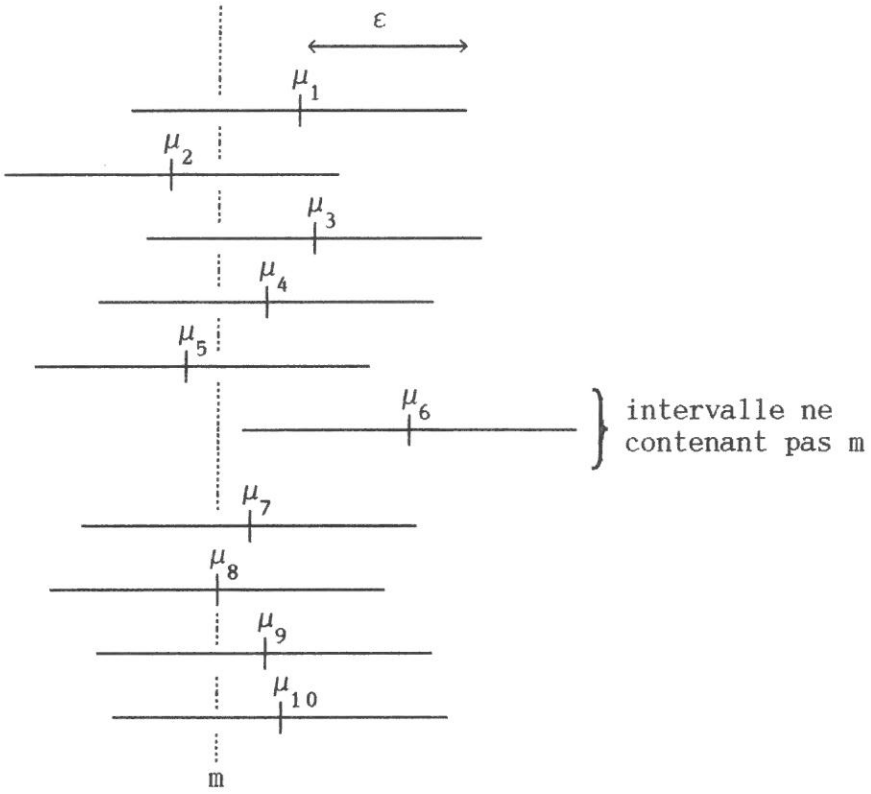
On retient alors comme estimation ponctuelle de la variance s^2 de la population la valeur observée de $\frac{n}{n-1}\ V$ sur $\mathcal{E}$; c'est la <u>variance corrigée</u> : $\sigma^2 = \frac{n}{n-1} \times \sigma_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$.

Remarque : Posons $S = \sqrt{\frac{n}{n-1}\ V}$; si S n'est pas constante, on n'a pas $E(S) = s$ (on a seulement $E(S) < s$), S n'est pas un estimateur sans biais de s.

II - Estimation d'une moyenne par intervalle de confiance

On veut estimer la moyenne m d'un caractère quantitatif dans une population $\mathcal{P}$. Une estimation à partir d'un échantillon donné ne renseigne pas beaucoup sur le degré d'approximation de m. On cherche, donc, un "intervalle aléatoire" dépendant de l'échantillon prélevé, tel que la probabilité qu'il contienne m soit acceptable. Cette probabilité p sera appelée "<u>niveau de confiance</u>" de l'estimation.

Pour mieux comprendre, prenons un niveau de confiance de 90 %. A chaque échantillon de moyenne μ, on associe un intervalle centré en μ : $[\mu - \varepsilon ; \mu + \varepsilon]$ où ε est choisi de sorte que la répartition sur 10 échantillons se fasse en moyenne de la façon suivante :



On considère la variable aléatoire $\bar{X}$ définie sur l'ensemble des échantillons aléatoires de taille n et qui prend pour valeur la moyenne du caractère dans les échantillons.

La méthode que nous allons développer concerne les cas suivants pour la distribution du caractère étudié dans la population :

- **la distribution est normale d'écart type connu s** : $\bar{X}$ suit la loi normale $\mathcal{N}\left(m ; \frac{s}{\sqrt{n}}\right)$; $T = \frac{\bar{X} - m}{s/\sqrt{n}}$ suit la loi normale $\mathcal{N}(0 ; 1)$

- **la distribution est normale d'écart type inconnu s** : $T = \frac{\bar{X} - m}{S/\sqrt{n}}$ suit la loi de Student à $n - 1$ degrés de liberté que l'on peut approcher par la loi normale $\mathcal{N}(0 ; 1)$ si l'échantillon est de grande taille (dans la pratique $n \geqslant 30$).

Exemple : Un grossiste achète un lot de plusieurs milliers de poulets à une coopérative agricole. Il voudrait estimer le poids moyen m de ces poulets à l'aide d'une "fourchette".

Plus précisément : Il voudrait un encadrement de m dont il serait "sûr" à 90 % (le niveau de confiance est 0,9). Le nombre $\alpha = 1 - 0{,}9 = 0{,}1$ (ou $\alpha = 10$ %) est appelé seuil de confiance. Pour cela, il prélève au hasard n poulets du lots. Il peut alors estimer ponctuellement m par la moyenne μ de l'échantillon et l'écart type du poids des poulets s par l'écart type corrigé σ de l'échantillon.

On cherche donc une valeur x telle que $P(\bar{X} \in [m-x\,;\, m+x]) = 0{,}9$.

La loi de $T = \dfrac{\bar{X} - m}{S/\sqrt{n}}$ peut être approchée par la loi normale $\mathcal{N}(0\,;\,1)$, on se ramène à chercher une valeur t telle que $P(-t \leqslant T \leqslant t) = 0{,}9$ soit $2\,\pi(t) - 1 = 0{,}9$ où π désigne la fonction de répartition de la loi normale centrée réduite ; d'où $\pi(t) = 0{,}95$. La lecture de la table de la loi normale centrée réduite donne $t = 1{,}64$.

D'où $P\left(-1{,}64 \leqslant \dfrac{\bar{X} - m}{S/\sqrt{n}} \leqslant 1{,}64\right) = 0{,}90$, une valeur observée de $\dfrac{\bar{X} - m}{S/\sqrt{n}}$ est dans l'intervalle de centre 0 et de rayon 1,64 avec la probabilité 0,9.

On a encore $P\left(\bar{X} - 1{,}64\,\dfrac{S}{\sqrt{n}} \leqslant m \leqslant \bar{X} + 1{,}64\,\dfrac{S}{\sqrt{n}}\right) = 0{,}9$.

En remplaçant $\bar{X}$ et S par leurs valeurs observées μ et σ, on obtient, à partir de l'observation d'un échantillon, l'intervalle $\left[\mu - 1{,}64\,\dfrac{\sigma}{\sqrt{n}}\ ;\ \mu + 1{,}64\,\dfrac{\sigma}{\sqrt{n}}\right]$ appelé intervalle de confiance de m au niveau de confiance de 90 % ou au seuil de confiance de 10 %.

Application numérique : $n = 60$; $\mu = 1{,}5$; $\sigma = 0{,}2$.

Vérifier que [1,45 ; 1,55] est un intervalle de confiance de m au niveau de confiance de 90 %.

Exercice 3 :

En reprenant les mêmes données, déterminer un intervalle de confiance de la moyenne m au seuil de 1 %.

On trouvera : [1,43 ; 1,57]. Commentaire ?

Exercice 4 :

En reprenant l'exercice 1 sur les paquets de beurre, déterminer un intervalle de confiance au seuil de 2 % pour la masse des paquets de beurre.
(On trouvera : [249,6 ; 250,8].)

Généralisation :

On considère dans une population $\mathcal{P}$, un échantillon $\mathcal{E}$ de taille n.
Soit α un seuil de confiance, $1-\alpha$ le niveau de confiance associé.

Soit t_α tel que $P(-t_\alpha \leqslant T \leqslant t_\alpha) = 1-\alpha$ (où T suit la loi $\mathcal{N}(0\,;\,1)$). On a alors : $\pi(t_\alpha) = 1 - \frac{\alpha}{2}$.

Soit μ et σ les valeurs observées de $\bar{X}$ et S sur l'échantillon $\mathcal{E}$; $\left[\mu - t_\alpha \frac{\sigma}{\sqrt{n}}\ ;\ \mu + t_\alpha \frac{\sigma}{\sqrt{n}}\right]$ est un intervalle de confiance de la moyenne m au seuil de confiance α.

Exercice 5 : (Extrait B.T.S. Analyse Biologique 1984)

On reprend les données de l'exercice 2. On suppose que la variable aléatoire $\bar{X}$ qui, à tout échantillon de taille $n = 100$, associe la glycémie moyenne de cet échantillon suit la loi normale $\mathcal{N}\left(m\,;\,\frac{s}{\sqrt{n}}\right)$ et on prend pour valeur de s l'estimation ponctuelle obtenue.

a) - Déterminer un intervalle de confiance de la glycémie moyenne m dans la population avec le niveau de confiance de 99 %.

b) - Quelle devrait être la taille de l'échantillon pour connaître avec le niveau de confiance de 95 % la glycémie moyenne m dans la population à 0,5 mg/100 ml près ?

Corrigé :

a) - Les valeurs observées de $\bar{X}$ et S sont $\mu = 92,20$ et $\sigma = 7,21$.
Pour $\alpha = 1$ %, $t_\alpha = 2,57$ et l'intervalle de confiance obtenu au vu de l'échantillon est : [90,34 ; 94,06].

b) - Un intervalle de confiance a pour bornes : $\bar{X} - t_\alpha \frac{s}{\sqrt{n}}$ et $\bar{X} + t_\alpha \frac{s}{\sqrt{n}}$. Comme on ne connaît pas s, on le remplace par son estimation ponctuelle $\sigma = 7,21$.
Pour connaître la glycémie moyenne à 0,5 mg/100 ml près, il faut donc que $t_\alpha \frac{\sigma}{\sqrt{n}} \leqslant 0,5$; soit $2\, t_\alpha\, \sigma \leqslant \sqrt{n}$ d'où $4\, t_\alpha^2\, \sigma^2 \leqslant n$.
Ici, le seuil de confiance est 5 %, $t_\alpha = 1,96$; on obtient $n \geqslant 799$.

Exercice 6 : (Extrait B.T.S. Construction Navale 1985)

On a mesuré les longueurs en mm d'un échantillon de 100 tiges d'acier, tirées au hasard, à la sortie d'une machine automatique. On a obtenu les résultats suivants en regroupant par classe les mesures :

longueur	[132;134[	[134;136[	[136;138[	[138;140[	[140;142[
effectif	2	5	13	24	19

longueur	[142;144[	[144;146[	[146;148[	[148;150[	[150;152[
effectif	14	10	8	3	2

a) - Calculer moyenne et écart type des longueurs des tiges de cet échantillon.

b) - A partir des résultats obtenus pour cet échantillon, proposer une estimation ponctuelle de la moyenne m et de l'écart type s de la longueur des tiges produites par cette machine.

c) - En supposant que la variable aléatoire qui, à tout échantillon de 100 tiges associe la moyenne de la longueur des tiges de l'échantillon, suit la loi normale $\mathcal{N}\left(m ; \frac{s}{\sqrt{n}}\right)$, donner un intervalle de confiance à 99 % de la longueur moyenne m des tiges de toute la production de la machine.

d) - Quelle doit être la taille d'un échantillon extrait de la production pour que la moyenne des longueurs des tiges soit estimée à 10^{-1} près, avec 95 % de certitude ?

Corrigé :

a) - La calculatrice donne 141,14 comme moyenne et 3,88 comme écart type.

b) $\mu = 141{,}14 \quad \sigma = 3{,}88 \sqrt{\frac{100}{99}} \approx 3{,}90$. **c)** - [140,13 ; 142,15]. **d)** $n \geq 5\ 844$.

Exercice 7 : (Extrait B.T.S. Agricole)

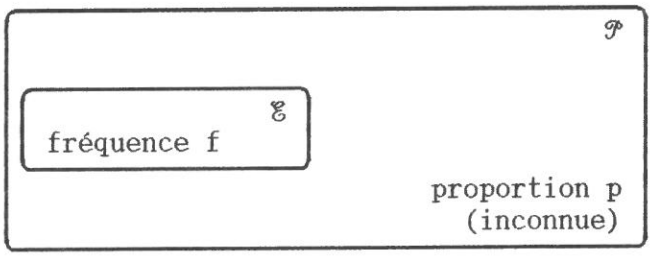
On veut contrôler la fabrication d'un fromage obtenu en grande série. On s'intéresse au poids des fromages sortant de la fabrication et l'on note X la variable aléatoire prenant pour valeur la masse des fromages (exprimée en grammes) ; on admet que X suit une loi normale. Un échantillon aléatoire de 51 fromages pesés unité par unité donne une moyenne de 263 g et un écart type de 7 g.

On demande de faire une estimation par intervalle de confiance au seuil de 5 % de la moyenne de X.

Corrigé : On estime s l'écart type de X par $\sigma = 7 \sqrt{\frac{51}{50}} \approx 7{,}07$. On obtient alors [261 ; 265] comme intervalle de confiance.

B - Travail sur proportion et fréquence

I - Estimation ponctuelle



Soit $\mathcal{P}$ une population dans laquelle on trouve une proportion p d'individus ayant un caractère A.

Soit $\mathcal{E}$ un échantillon aléatoire de $\mathcal{P}$, de taille n, où f est la fréquence de ce caractère : f est une estimation ponctuelle de p.

Exemple :

Un sondage effectué auprès de 150 personnes choisies de façon aléatoire dans une circonscription donne 45 suffrages au candidat A.

On a $f = 0,3$.

On choisit comme estimation ponctuelle de la proportion p d'électeurs favorables au candidat A, ce nombre 0,3.

Pourquoi cette estimation ponctuelle est-elle une bonne estimation ?

Soit X et F les variables aléatoires définies sur l'ensemble des échantillons de taille n prenant pour valeurs respectives le nombre d'individus possédant le caractère et la fréquence du caractère A ; X suit la loi binomiale $\mathcal{B}(n ; p)$ et $F = \frac{X}{n}$. Ainsi, en posant $q = 1-p$, on obtient : $E(X) = np$, $Var(X) = npq$, $E(F) = p$ et $Var(F) = \frac{pq}{n}$.

Comme $E(F) = p$, F est un estimateur sans biais de p et comme de plus $\lim_{n \to +\infty} Var(F) = 0$, F est un estimateur convergent de p.

II - Estimation par intervalle de confiance

On cherche un intervalle de confiance de p au seuil α. On considère la variable aléatoire F définie sur l'ensemble des échantillons de taille n et qui prend pour valeur la fréquence du caractère dans les échantillons.

La loi de $T = \dfrac{F - p}{\sqrt{\dfrac{F(1-F)}{n}}}$ peut être approchée par la loi normale $\mathcal{N}(0 ; 1)$ pour n assez grand ($n \geqslant 30$).

Exemple :

On veut estimer le nombre N d'oiseaux d'une certaine espèce dans une région. Pour cela, on en capture une centaine que l'on bague, puis que l'on relâche. Quelque temps après, on en capture à nouveau 100 : après chaque capture, on observe si l'animal est bagué ou non, puis on le relâche (tirage avec remise). Le nombre d'oiseaux bagués ainsi observé est 17.

Soit $p = \frac{100}{N}$ le pourcentage d'oiseaux capturés. La seconde capture s'assimile à un échantillon de taille 100 de la population et $f = \frac{17}{100}$ est une estimation ponctuelle de p.

Nous allons construire un intervalle de confiance de p. Pour cela fixons d'abord un niveau de confiance de 95 %. ($\alpha = 0,05$).
La taille de l'échantillon étant suffisamment importante, la loi de $T = \frac{F - p}{\sqrt{\frac{F(1-F)}{100}}}$ est approchée par la loi normale $\mathcal{N}(0 ; 1)$.

On a donc $P(-t \leqslant T \leqslant t) = 0,95$, donc $\pi(t) = 0,975$, soit $t = 1,96$.

Ainsi $$P\left(-1,96 \leqslant \frac{F - p}{\sqrt{\frac{F(1-F)}{100}}} \leqslant 1,96\right) = 0,95.$$

D'où $$P\left(F - 1,96\sqrt{\frac{F(1-F)}{100}} \leqslant p \leqslant F + 1,96\sqrt{\frac{F(1-F)}{100}}\right) = 0,95.$$

En remplaçant alors F par sa valeur observée f, on obtient $\left[f - 1,96\sqrt{\frac{f(1-f)}{100}} \; ; \; f + 1,96\sqrt{\frac{f(1-f)}{100}}\right]$ comme intervalle de confiance de p au niveau de confiance de 95 %.

Application numérique :

$f = 0,17$, on obtient l'intervalle $[0,09;0,25]$.

Donc $0,09 \leqslant p \leqslant 0,25$; d'où $0,09 \leqslant \frac{100}{N} \leqslant 0,25$.

Ainsi, $400 \leqslant N \leqslant 1\,112$ (au seuil de confiance de 5 %).

Généralisation

On considère dans une population $\mathcal{P}$, un échantillon $\mathcal{E}$ de taille n.
Soit α un seuil de confiance, $1-\alpha$ le niveau de confiance associé.
Soit t_α tel que $P(-t_\alpha \leqslant T \leqslant t_\alpha) = 1-\alpha$ (où T suit la loi $\mathcal{N}(0\,;\,1)$), on a

$$\pi(t_\alpha) = 1 - \frac{\alpha}{2}.$$

Soit f la valeur observée de F dans l'échantillon $\mathcal{E}$; on obtient alors $\left[f - t_\alpha \sqrt{\frac{f(1-f)}{n}} \;;\; f + t_\alpha \sqrt{\frac{f(1-f)}{n}}\right]$ comme intervalle de confiance de la proportion p du caractère A dans la population $\mathcal{P}$ au seuil de α.

Exemple :

Un centre de transfusion sanguine désire connaître, à 0,05 près, la proportion p de personnes du groupe sanguin 0 (donneurs universels) dans sa zone d'action et cela au niveau de confiance de 99 %.
Déterminer la taille de l'échantillon à prélever dans cette population pour satisfaire cette demande.

Avec les mêmes notations que précédemment: $\alpha = 0{,}01$ et $t_\alpha = 2{,}57$.

Un intervalle de confiance de p, au seuil de 1 %, est :

$$\left[f - 2{,}57 \sqrt{\frac{f(1-f)}{n}} \;;\; f + 2{,}57 \sqrt{\frac{f(1-f)}{n}}\right].$$

Si l'on veut connaître p à 5 %, il faut donc choisir n de sorte que $|p - f| < 0{,}05$, soit $2{,}57 \sqrt{\frac{f(1-f)}{n}} < 0{,}05$.

Majorons $\sqrt{f(1-f)}$ par $\frac{1}{2}$, on obtient $\frac{1{,}285}{\sqrt{n}} < 0{,}05$ soit $n \geqslant 661$.

Remarque :

La proportion d'individus du groupe 0 dans la population française est environ $\frac{1}{3}$. En remplaçant p par $\frac{1}{3}$ dans $\sqrt{\frac{f(1-f)}{n}}$, on obtient $n \geqslant 588$ et on reste donc dans le même ordre de grandeur.

Exercice 8 :

Une semaine avant des élections, un institut de sondage a interrogé, au hasard, n personnes (n est de l'ordre de plusieurs centaines) sur leurs intentions de vote.

L'institut donne comme intervalle de confiance au seuil de 5 %, la fourchette 34,72 % - 43,37 % pour le pourcentage d'électeurs favorables au candidat Dupont.

Déterminer la taille n de l'échantillon interrogé par l'institut de sondage.

Corrigé :

L'intervalle de confiance de la proportion d'électeurs favorables au candidat Dupont est $\left[f - t_\alpha \sqrt{\frac{f(1-f)}{n}} \; ; \; f + t_\alpha \sqrt{\frac{f(1-f)}{n}}\right]$ où f est la fréquence observée dans l'échantillon et n la taille de l'échantillon.

On a alors $f = \frac{0,4337 + 0,3472}{2} = 0,39045$ et $2\, t_\alpha \sqrt{\frac{f(1-f)}{n}} = 0,4337 - 0,3472$,

ainsi $2\, t_\alpha \sqrt{\frac{f(1-f)}{n}} = 0,0865$ et $n = \frac{4\, f(1-f)}{(0,0865)^2} \times t_\alpha^2$.

Au seuil de $\alpha = 5$ %, $t_{0,05} = 1,96$; $n = \frac{4 \times 0,39045(1 - 0,39045)}{(0,0865)^2} \times 1,96^2 \approx 489$.

Tests de validité d'hypothèses

On a souvent à établir des lois générales à partir de l'observation de quelques cas particuliers, donc à prendre des décisions à partir de l'étude d'un échantillon. Cela est valable si la partie observée est, relativement aux critères observés, semblable au tout. Mais, ce procédé est entâché de risques d'erreur qu'il faut tenter de mesurer.
C'est, par exemple, la méthode employée pour la décision de mise sur le marché d'un médicament, pour le contrôle de qualité de fabrication, pour l'étude de l'efficacité d'une campagne publicitaire, pour la comparaison de rendement de deux variétés de céréales...

A - Généralités sur les tests de validité d'hypothèse

1°) - Définition

Un test de validité d'hypothèse est un mécanisme permettant de confirmer ou d'infirmer une hypothèse par l'observation d'un échantillon, c'est donc une fonction définie de l'ensemble des échantillons vers l'ensemble des décisions $\{d_0 ; d_1\}$ où d_0 et d_1 sont respectivement l'acceptation et le refus de l'hypothèse.

2°) - Exemples

a) - Commentaire d'un test élaboré

Pour procéder à l'embauche d'un graphologue au service du personnel, le chef du personnel d'une grosse entreprise envisage un test. Il propose 12 paires d'écritures constituées de l'écriture d'un médecin et de celle d'un avocat ; le candidat sera embauché s'il identifie, pour au moins 9 paires, l'écriture du médecin et l'écriture de l'avocat.

Deux types de risque d'erreur

- du point de vue du chef du personnel :

L'engagement d'un incompétent étant pour l'entreprise beaucoup plus grave que le rejet d'un candidat de valeur, le chef du personnel veut déterminer la probabilité d'embaucher un incompétent.
Si le candidat répond au hasard, la probabilité p qu'il reconnaisse une paire d'écritures donnée est 0,5. La variable aléatoire T égale au nombre de bonnes réponses suit alors la loi binomiale de paramètres 12 et 0,5.

Soit α la probabilité cherchée : $\alpha = P(T \geqslant 9) = \sum_{n=9}^{n=12} C_{12}^{n} \left(\frac{1}{2}\right)^{12}$, d'où $\alpha \approx 7{,}3$ %. Ce que l'on peut représenter ainsi (**figure 1**) :

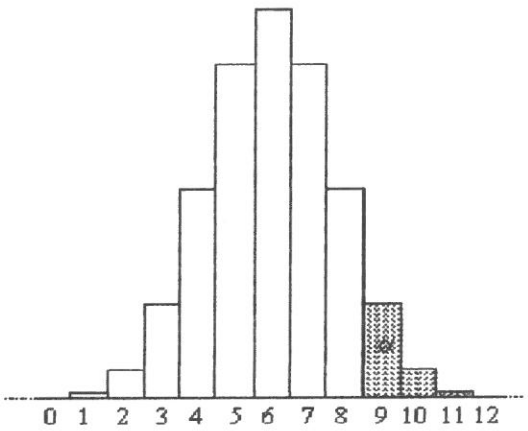


figure 1

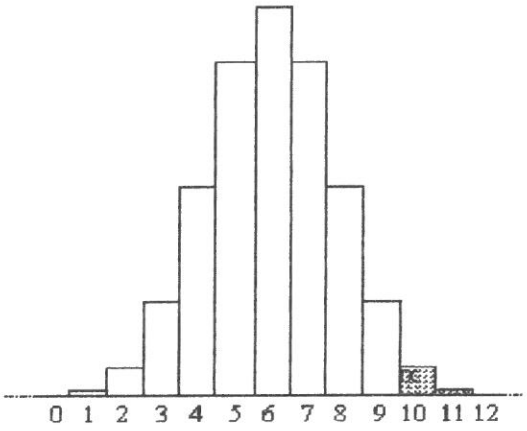


figure 2

(Chaque rectangle représente, par son aire, la probabilité que T prenne la valeur indiquée à sa base.)

Cherchant à réduire cette probabilité, le chef du personnel pense exiger 10 bonnes réponses ; alors α devient : $P(T \geqslant 10) = \sum_{n=10}^{n=12} C_{12}^{n} \left(\frac{1}{2}\right)^{12}$, d'où $\alpha \approx 1{,}93$ % (**figure 2**). Cela semble trop sévère

En prenant 13 paires d'écritures et en exigeant 10 bonnes réponses, T suit maintenant la loi binomiale de paramètres 13 et 0,5 : $\alpha = P(T \geqslant 10) = \sum_{n=10}^{n=13} C_{13}^{n} \left(\frac{1}{2}\right)^{13} \approx 4{,}6$ %, cela raisonnable (**figure 3**).

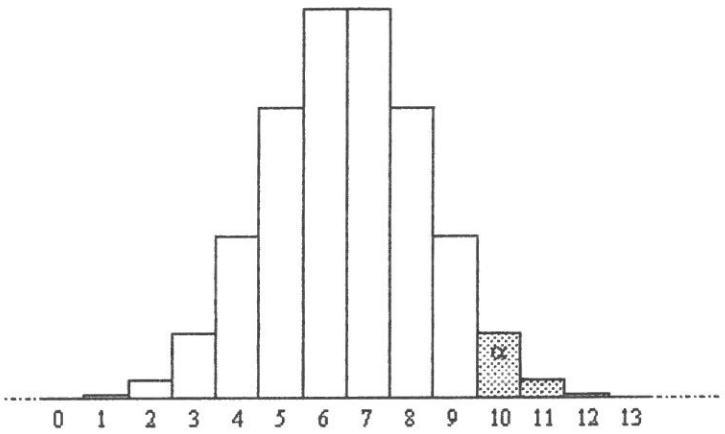


figure 3

- du point de vue du candidat :

Plaçons nous maintenant du point de vue d'un candidat qui estime à 85 % la probabilité qu'il reconnaisse une paire d'écritures donnée.

T suit alors la loi binomiale de paramètres 13 et 0,85. La probabilité qu'il fournisse moins de 10 bonnes réponses sur 13 est (**figure 4**) :

$$\beta = P(T \leqslant 9) = 1 - P(T \geqslant 10) = 1 - \sum_{n=10}^{n=13} C_{13}^{n}\, 0{,}85^{n} \times 0{,}15^{13-n} \approx 11{,}8\ \%.$$

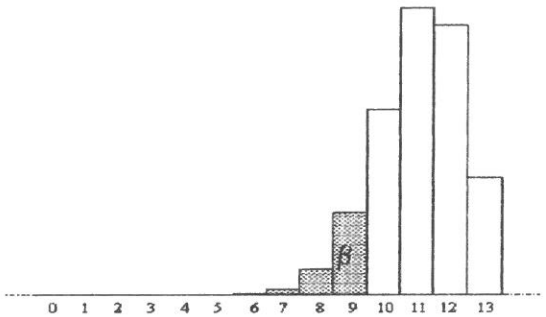


figure 4

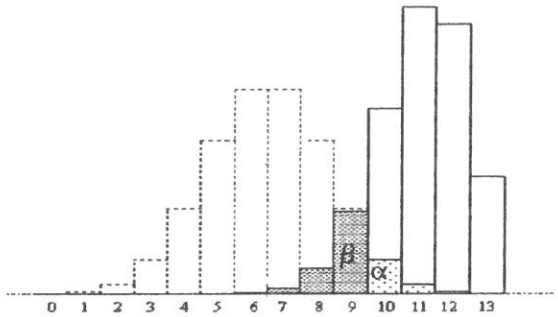


figure 5

La **figure 5** met en évidence les variations simultanées de α et β.

b) - Construction d'un test

On considère des tubes électriques d'un même type provenant de deux fabriques. La durée de vie des tubes est une variable aléatoire qui suit une loi normale. On cherche à savoir si les deux usines produisent des tubes identiques.

Dans la première production, d'écart type 24, on prélève un échantillon représentatif de taille 100 pour lequel la durée de vie moyenne des tubes est 1 452 heures. Dans la seconde production d'écart type 28, on prélève un échantillon représentatif de taille 200 pour lequel la durée de vie moyenne des tubes est 1 447 heures.

On suppose que les deux usines produisent des tubes identiques alors les durées de vie moyennes des tubes des deux productions sont les mêmes.

On peut considérer que les durées de vie moyennes des tubes dans les deux échantillons sont les réalisations de deux variables aléatoires indépendantes X_1 et X_2. D'après les résultats sur les lois d'échantillonnage, la variable aléatoire $T = \dfrac{X_1 - X_2}{\sqrt{\dfrac{24^2}{100} + \dfrac{28^2}{200}}}$ suit la loi normale $\mathcal{N}(0\ ;\ 1)$.

On cherche un intervalle I centré en 0 qui contienne les valeurs observées de T avec une probabilité de 95 %. En utilisant une table de loi normale centrée, réduite, on a : $P[|T| < 1,96] = 0,95$ donc $I =]-1,96 ; 1,96[$.

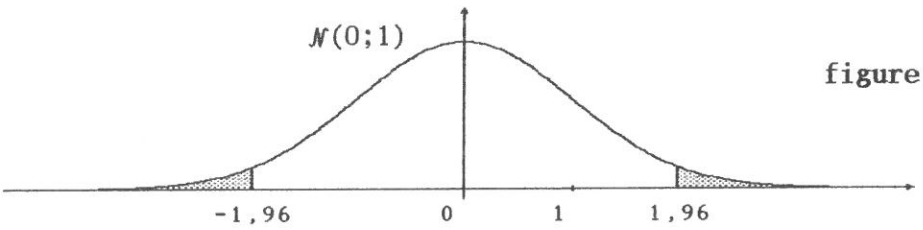


figure 6

On considére que l'événement "$|T| > 1,96$" est rare et on pense que, s'il est réalisé lors de l'observation d'un échantillon, l'hypothèse de départ est fausse. Ici, T prend la valeur $T_{obs} = 1,67$; comme $|T_{obs}| < 1,96$, on peut accepter le fait que les usines fabriquent des tubes identiques.

3°) - A propos de la méthode : analogie avec un schéma d'urne

On est en présence d'une urne contenant des boules de trois couleurs : bleue (B), rouge (R) et verte (V). On ne connaît pas la répartition de ces trois couleurs et on souhaite savoir si elle est du type peu de vertes, peu de rouges et beaucoup de blanches au vu du tirage d'une seule boule.

- **on sait que l'urne contient peu de verte :**
  - * si on tire une boule blanche ou verte (ce qui est peu probable), on accepte l'hypothèse de répartition ;
  - * si on tire une boule rouge, on refuse l'hypothèse de répartition aux dépens d'une proportion de boules rouges plus importante.
- **si on ne dispose pas de renseignement supplémentaire :**
  - * si on tire une boule blanche, on accepte l'hypothèse ;
  - * si on tire une boule verte ou rouge : on refuse l'hypothèse aux dépens d'une répartition plus importante en boules vertes ou en boules rouges selon la couleur de la boule tirée.

Ici les boules sont assimilables aux échantillons et leurs couleurs représentent les zones de valeurs observables du paramètre testé.

4°) - Elaboration d'un test

a) - Formulation des hypothèses

On émet deux hypothèses H_0 et H_1 :

H_0 est l'hypothèse nulle, c'est l'hypothèse que l'observateur croit vraie et qu'il ne rejettera que si elle est vraiment infirmée par l'expérience. C'est sous cette hypothèse que s'effectue le calcul des probabilités.

C'est en général une égalité portant sur un paramètre de la population (moyenne ou proportion dans les programmes de S.T.S.), elle correspond à : - une hypothèse facile à formuler,
- une hypothèse de stabilité,
- une hypothèse de prudence, de bon sens.

H_1 est l'hypothèse alternative , c'est une hypothèse que l'on estime a priori peu probable mais possible. Sa forme dépend du contexte, c'est en général : - une non égalité et le test est dit bilatéral,
- une inégalité ou une égalité et le test est dit unilatéral.

Quand H_1 n'est pas le contraire de H_0, le test est dit non-contradictoire.

b) - Détermination de la variable de décision

C'est une variable aléatoire T dont on exprime la loi sous l'hypothèse H_0 à partir des résultats sur l'échantillonnage. Elle s'exprime à partir de la loi suivie par le paramètre des échantillons.

c) - Seuil de signification

On fixe α le seuil de signification ou de confiance du test. α est la probabilité que l'hypothèse H_0 soit rejetée alors qu'elle est vraie.

Le choix de α, en général 0,01 ou 0,05 ou 0,1 est lié à la confiance que l'on accorde à l'hypothèse H_0 : si l'on est persuadé que H_0 est vraie, l'observation d'un événement infirmant H_0 ne doit être que très rare, on choisira une valeur de α petite ($\alpha = 0{,}01$) ; si au contraire, H_0 paraît hasardeuse, on choisira une valeur de α plus grande ($\alpha = 0{,}1$).
Lorsque $\alpha = 0{,}05$, on dit que le test est significatif. Lorsque $\alpha = 0{,}01$, on dit que le test est très significatif.
Puis, on détermine la région d'acceptation. C'est un intervalle I_α, en général de la forme [a,b] ou $]-\infty$, b] ou [a,$+\infty$[, qui dépend de T, de H_0, de H_1 et de α et qui vérifie $P[T \in I_\alpha] = 1 - \alpha$.
La région critique ou région de rejet est le complémentaire de I_α dans $\mathbb{R}$.
$1 - \alpha$ est le niveau de confiance du test.

d) - Règle de décision

Sous l'hypothèse H_0, le fait que T prenne des valeurs hors de I_α est relativement rare. Appelons T_{obs} la valeur prise par T pour l'échantillon observé.

La règle de décision est la suivante :

- si $T_{obs} \notin I_\alpha$: on n'admet pas l'effet du hasard dans le choix de l'échantillon et on rejette, dans ce cas, l'hypothèse H_0,
- si $T_{obs} \in I_\alpha$: on accepte l'hypothèse H_0.

Remarque : Si T_{obs} est proche des bornes de I_α, il vaut mieux renouveler le test avec un échantillon de taille supérieure, mais cela entraîne une augmentation du coût du test.

e) - Mise en oeuvre du test

On met en place un protocole expérimental pour prélever un échantillon et recueillir les observations de la façon la plus fiable possible. On calcule T_{obs} et on applique la règle de décision.

5°) - A propos des risques d'erreur

Naturellement, comme on ne dispose pas de renseignements sur l'ensemble de la population, on risque de se tromper en prenant la décision et il importe de contrôler au maximum tout risque d'erreur.

A l'issue d'un test, différents cas de figure peuvent se présenter :

- on accepte avec raison l'hypothèse nulle,
- on rejette avec raison l'hypothèse nulle,
- on accepte à tort l'hypothèse nulle,
- on rejette à tort l'hypothèse nulle.

Examinons les probabilités de chacun des deux derniers cas qui correspondent à une mauvaise décision :

- la probabilité de rejeter H_0 à tort est le risque d'erreur de première espèce, c'est α le seuil de signification du test ; cette probabilité est contrôlée par le statisticien,
- la probabilité β d'accepter H_0 à tort est le risque d'erreur de seconde espèce. β ne peut être calculé qu'en faisant une hypothèse sur la valeur du paramètre. β doit être le plus petit possible. La puissance d'un test est $1-\beta$.

En résumé, on a le tableau de probabilités suivant :

		vérité	
		H_0 est vraie	H_0 est fausse
décision	Acceptation de H_0	$1-\alpha$	β
	Rejet de H_0	α	$1-\beta$

B - Exemples

I - Test de validité d'hypothèse relatif à un paramètre

1°) - Test de validité d'hypothèse relatif à une fréquence

Dans une population, on étudie la fréquence p d'un caractère quantitatif. Les hypothèses concernent p.

La fréquence du caractère dans un échantillon de taille n peut être considérée comme la réalisation d'une variable aléatoire X.

On prend H_0 : "$p = p_0$" comme hypothèse nulle où p_0 est une valeur donnée.

La variable de décision est $T = \dfrac{X - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$ dont on approche la loi, sous l'hypothèse H_0, par la loi normale $\mathcal{N}(0\,;\,1)$ si $n \geqslant 30$ et $np_0 \geqslant 15$.

L'hypothèse alternative H_1 s'exprime, en général, sous l'une des formes suivantes : "$p \neq p_0$" , "$p < p_0$" , "$p > p_0$" ou "$p = p_1$" (p_1 étant une valeur donnée). La première forme conduit à un test bilatéral, les autres à des tests unilatéraux.

Situation 1

On considère un jeu de 52 cartes composé de 26 rouges et 26 noires. Sans les voir, une personne identifie la couleur de 33 cartes sur 52. Peut-on affirmer que la personne est extrasensorielle au seuil de 5 % ? au seuil de 1 % ?

Corrigé

On teste une hypothèse relative à un pourcentage. L'hypothèse nulle est H_0 : "la personne répond au hasard" qui peut se mettre sous la forme : "$p = 0,5$" où p désigne le probabilité que la personne reconnaisse la couleur d'une carte donnée.

Le pourcentage de bonnes réponses dans l'échantillon de taille 52 peut être considéré comme la réalisation d'une variable aléatoire X. La variable de décision est $T = \dfrac{X - 0,5}{\sqrt{\dfrac{0,25}{52}}}$ dont on approche la loi, sous l'hypothèse H_0, par la loi normale $\mathcal{N}(0\,;\,1)$ car $52 \geqslant 30$ et $52 \times 0,5 = 26 \geqslant 15$.

On teste H_0 : "$p = 0,5$" contre H_1 : "$p > 0,5$"

("la personne répond au hasard" contre "la personne est extrasensorielle").

Le test est unilatéral. Soit α le seuil de signification du test.

Sous l'hypothèse H_0, le fait d'avoir un pourcentage de bonnes réponses "très supérieur" à 1/2 est rare ; la valeur critique est le réel t_α tel que $P[T < t_\alpha] = 1 - \alpha$. La règle de décision est :

- si $T_{obs} > t_\alpha$: on n'admet pas l'effet du hasard dans le choix de l'échantillon et on refuse H_0 ;
- si $T_{obs} < t_\alpha$: on accepte H_0.

Ici, $T_{obs} = 1{,}94$.

On prend $\alpha = 5\ \%$: $t_{0,05} = 1{,}64$. $1{,}94 > t_{0,05}$: on refuse H_0.

La personne a peut-être un pouvoir, au risque d'erreur de 5 %.

On prend $\alpha = 1\ \%$: $t_{0,01} = 2{,}33$. $1{,}94 < t_{0,01}$: on accepte H_0.

La personne a peut-être répondu au hasard au seuil de confiance de 1 %.

Situation 2

Jusqu'à présent, pour soigner une maladie, on ne connaissait qu'un médicament A qui guérissait à 45 %. On teste un nouveau médicament B sur 50 personnes.

i) - Elaborer un test de H_0 : "B est aussi efficace que A" contre H_1 : "B est plus efficace que A" au seuil de 5 %.

ii) - Le fabricant sait que son médicament est efficace à 65 %. Calculer le risque de seconde espèce.

Corrigé

i) - On teste une hypothèse relative à un pourcentage.

L'hypothèse nulle est H_0 : "$p = 0{,}45$" où p désigne la proportion de guérisons avec le médicament B.

La proportion p de guérisons dans l'échantillon peut être considérée comme la réalisation d'une variable aléatoire X.

La variable de décision est $T = \dfrac{X - 0{,}45}{\sqrt{\dfrac{0{,}45 \times 0{,}55}{50}}}$ dont la loi est approchée, sous l'hypothèse H_0, par la loi normale $\mathcal{N}(0\ ;\ 1)$ car $50 \geqslant 30$ et $50 \times 0{,}45 \geqslant 15$.

(La loi de X est approchée par la loi normale $\mathcal{N}(0{,}45\ ;\ 0{,}0703)$).

On teste H_0 : "p = 0,45" contre H_1 : "p > 0,45" ("le nouveau médicament est aussi efficace" contre "le nouveau médicament est plus efficace")

Le test est unilatéral. Le seuil de confiance est 5 %.

Sous l'hypothèse H_0, le fait d'avoir un pourcentage de guérisons "très supérieur" à 0,45 est rare ; la valeur critique est le réel $t_{0,05}$ tel que $P[T < t_{0,05}] = 0,95$.

On trouve $t_{0,05} = 1,64$, finalement :

- si $T_{obs} > 1,64$: on n'admet pas l'effet du hasard dans le choix de l'échantillon et on refuse H_0,
- si $T_{obs} < 1,64$: on accepte H_0.

ii) - Détermination du risque de seconde espèce β.

On est maintenant sous l'hypothèse "p = 0,65".

β est la probabilité d'accepter H_0, c'est la probabilité que le médicament B ne soit pas déclaré plus efficace que A lors du test précédent.

$\beta = P[T < 1,64]$ mais on ne connaît pas la loi de T ; on sait seulement que celle de $T' = \dfrac{X - 0,65}{\sqrt{\dfrac{0,35 \times 0,65}{50}}}$ est approchée par $\mathcal{N}(0 ; 1)$ (car $50 \geqslant 30$ et $50 \times 0,65 = 32,5 \geqslant 15$).

(La loi de X est approchée par la loi normale $\mathcal{N}(0,65 ; 0,067)$)

$$\beta = P\left[\frac{X - 0,45}{\sqrt{\dfrac{0,45 \times 0,55}{50}}} < 1,64\right] \approx P[X < 0,565]$$

d'où $$\beta \approx P\left[\frac{X - 0,65}{\sqrt{\dfrac{0,35 \times 0,65}{50}}} < \frac{0,565 - 0,65}{\sqrt{\dfrac{0,35 \times 0,65}{50}}}\right] \approx P[T' < -1,26] \approx 10,38\ \%.$$

Les résultats précédents sont illustrés dans la **figure 7**.

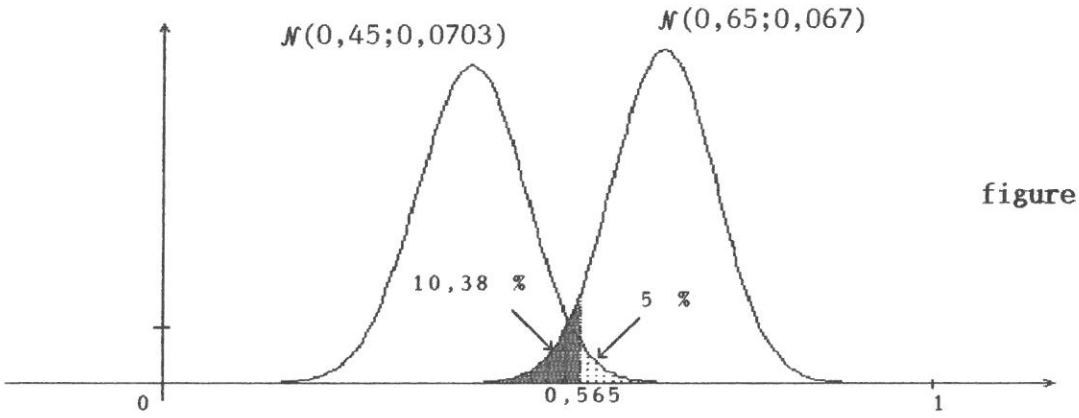


figure 7

2°) - Test de validité d'hypothèse relatif à une moyenne

Le paramètre étudié est la moyenne m d'un caractère quantitatif. On note s l'écart type du caractère. Les hypothèses concernent m.

On se place dans le cas d'échantillons non exhaustifs ou pouvant être considérés comme tels.

La moyenne du caractère dans un échantillon de taille n peut être considérée comme la réalisation d'une variable aléatoire X. On note V la variance de l'échantillon.

On prend H_0 : "$m = m_0$" comme hypothèse nulle. On se place sous cette hypothèse.

La variable de décision T prend différentes formes, selon les cas suivants :

Si s est connu : $T = \dfrac{X - m_0}{s \,/\, \sqrt{n}}$

a) - la loi de T est la loi normale $\mathcal{N}(0 ; 1)$ si le caractère est distribué normalement dans la population,

b) - la loi de T est approchée par la loi normale $\mathcal{N}(0 ; 1)$ si $n \geqslant 30$.

Si s est inconnu : $T = \dfrac{X - m_0}{\sigma \,/\, \sqrt{n}}$ où $\sigma^2 = \dfrac{n}{n-1} V$ est la variance empirique

c) - la loi de T est la loi de Student à $n - 1$ degrés de liberté si le caractère est distribué normalement dans la population ;

d) - la loi de T est approchée par la loi normale $\mathcal{N}(0 ; 1)$ si $n \geqslant 30$.

L'hypothèse alternative H_1 s'exprime, en général, sous l'une des formes suivantes : "$m \neq m_0$" ou "$m < m_0$" ou "$m > m_0$" ou "$m = m_1$" où m_1 est une valeur donnée.

Situation 3

Dans une banque, les dépôts suivent la loi $\mathcal{N}(200 ; 64)$. Dans un échantillon de taille 16, on observe une moyenne de 180.

i) - Cet échantillon est-il représentatif de la population ?

ii) - Cette moyenne observée est-elle anormalement plus faible que la moyenne dans la population ?

Corrigé

Les deux questions concernent des tests sur une moyenne.

L'hypothèse H_0 est "$m = 200$" où m désigne la moyenne des dépôts.

La moyenne du caractère dans l'échantillon peut être considérée comme la réalisation d'une variable aléatoire X.

On est dans le cas **a)** car les dépôts suivent une loi normale et on connaît l'écart type de la population $s = 64$.

La variable de décision est $T = \dfrac{X - 200}{64/\sqrt{16}} = \dfrac{X - 200}{16}$ qui suit, sous l'hypothèse H_0, la loi $\mathcal{N}(0 ; 1)$. La valeur observée de T est $T_{obs} = -1,25$.

i) - On teste H_0 : "$m = 200$" contre H_1 : "$m \neq 200$" ("la moyenne est 200" contre "la moyenne est différente de 200")

Le test est bilatéral. On prend $\alpha = 5\ \%$ pour seuil de confiance du test.

Sous l'hypothèse H_0, le fait d'avoir un échantillon de moyenne "très loin" de 200 est rare ; la valeur critique est le réel $t_{0,05}$ tel que $P[|T| < t_{0,05}] = 0,95$. On trouve $t_{0,05} = 1,96$.

$|T_{obs}| < 1,96$: on accepte H_0 et l'échantillon peut être considéré comme représentatif de la population au seuil de 5 %.

Remarque : Si T suit une loi symétrique par rapport à 0 et unimodale (la densité a un seul maximum), parmi tous les intervalles [a ; b] tels que $P[T \in [a ; b]] = 1 - \alpha$, celui où a et b sont opposés est de longueur minimale ce qui offre la meilleure précision.

ii) - On teste H_0 : "$m = 200$" contre H_1 : "$m < 200$" ("la moyenne est 200" contre "la moyenne est inférieure à 200")

Le test est unilatéral. On prend $\alpha = 5\ \%$ pour seuil de confiance du test.

Sous l'hypothèse H_0, le fait d'avoir un échantillon de moyenne "très inférieure" à 200 est rare ; la valeur critique est le réel $t'_{0,05}$ tel que $P[T > t'_{0,05}] = 0,95$. On trouve $t'_{0,05} = -1,64$.

$T_{obs} > -1,64$: on accepte H_0. Au seuil de 5 %, on peut dire que la moyenne observée dans l'échantillon n'est pas anormalement plus faible que la moyenne dans la population.

Situation 4

La durée de vie moyenne d'un échantillon de 100 tubes fluorescents a été établie par le calcul à 1 570 heures avec un écart type de 120 heures.

i) - Si m désigne la durée de vie moyenne des tubes de la production, tester l'hypothèse "m = 1 600" contre l'hypothèse "m ≠ 1 600", avec un seuil de confiance de 0,05 puis de 0,01

ii) - Sur un autre échantillon, la moyenne observée a été de 1 580 heures avec un écart type de 120 heures et au risque d'erreur de 5 % cet échantillon a été jugé représentatif de la population. Que peut-on dire de la taille de cet échantillon ?

Corrigé

i) - On teste une moyenne. L'hypothèse nulle est H_0 : "m = 1 600".

On est dans le cas **d)** car on ne connaît pas l'écart type de la population et que la taille de l'échantillon est supérieure à 30. La variable de décision est $T = \dfrac{X - 1\ 600}{120/\sqrt{99}}$ dont la loi est approchée, sous l'hypothèse H_0, par la loi normale $\mathcal{N}(0\ ;\ 1)$. La valeur observée de T est $T_{obs} = -2{,}5$.

On teste H_0 : "m = 1 600" contre H_1 : "m ≠ 1 600"

Le test est bilatéral.

Sous l'hypothèse H_0, le fait d'avoir un échantillon de moyenne "très loin" de 200 est rare ; la valeur critique est le réel t_α tel que $P[|T| < t_\alpha] = 1 - \alpha$.

Si $|T_{obs}| < t_\alpha$: on accepte H_0 et si $|T_{obs}| > t_\alpha$: on refuse H_0.

Si $\alpha = 5\ \%$: $t_{0,05} = 1{,}96$; $|T_{obs}| > t_{0,05}$, donc on refuse H_0.

Si $\alpha = 1\ \%$: $t_{0,01} = 2{,}576$; $|T_{obs}| < t_{0,01}$, on accepte H_0 mais le test est peu significatif car $|T_{obs}|$ est voisin de $t_{0,01}$.

ii) - Au risque de 5 %, l'échantillon a été jugé représentatif. On est dans les conditions du **a)**, $t_{0,05} = 1{,}96$.

On est dans la situation $|T_{obs}| < 1{,}96$, soit $\left|\dfrac{1\ 580 - 1\ 600}{120/\sqrt{n-1}}\right| < 1{,}96$

d'où $n \leqslant 139$. (Si la taille de l'échantillon avait été plus importante, il aurait fallu que α soit moindre pour que l'échantillon soit déclaré représentatif de la population.)

Remarque : L'augmentation de la taille de l'échantillon réduit l'écart type de la loi normale approchant la loi de X. La **figure 8** représente la densité de X dans les cas où la taille est 100 ou 500 ($\alpha = 5$ %).

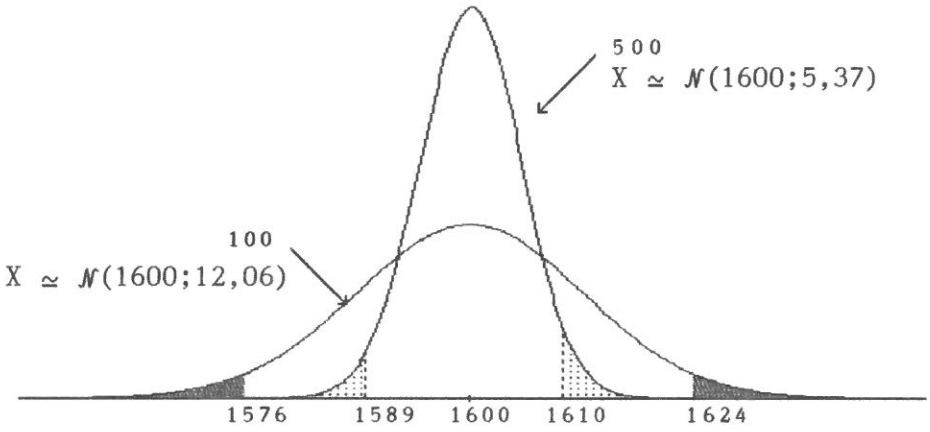


figure 8

II - Tests de comparaison de populations

1°) - Comparaison de deux moyennes

Dans deux populations $\mathcal{P}_1$ et $\mathcal{P}_2$, on étudie un caractère quantitatif ayant pour moyennes respectives m_1 et m_2 (inconnues) et pour écarts types respectifs s_1 et s_2.

Il s'agit de savoir, au vu d'un échantillon de chacune des deux populations, s'il existe une différence significative entre m_1 et m_2.

La moyenne du caractère dans un échantillon de taille n_i de la population $\mathcal{P}_i$ peut être considérée comme la réalisation d'une variable aléatoires X_i pour $i \in \{1 ; 2\}$. X_1 et X_2 sont indépendantes. On note σ_i l'écart type empirique de l'échantillon.

On élabore un test ; on prend comme hypothèse nulle H_0 : "$m_1 = m_2$".

On se place sous l'hypothèse H_0. La variable de décision T prend différentes formes, selon les cas suivants.

s_1 et s_2 sont connus : $T = \dfrac{X_1 - X_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$

a) - la loi de T est la loi normale $\mathcal{N}(0 ; 1)$ si le caractère est distribué normalement dans les deux populations,

b) - la loi de T est approchée par la loi normale $\mathcal{N}(0 ; 1)$ si n_1 et n_2 sont supérieurs ou égaux à 30.

s_1 et s_2 sont inconnus mais égaux :

$$T = \frac{X_1 - X_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \text{où} \quad \sigma^2 = \frac{(n_1 - 1)\ \sigma_1^2 + (n_2 - 1)\ \sigma_2^2}{n_1 + n_2 - 2}$$

c) - la loi de T est la loi de Student à $n_1 + n_2 - 2$ degrés de liberté si le caractère est distribué normalement dans les deux populations,

d) - la loi de T est approchée par la loi normale $\mathcal{N}(0\ ;\ 1)$ si le caractère est distribué normalement dans les deux populations et $n_1 + n_2 - 2 \geqslant 30$,

e) - la loi de T est approchée par la loi normale $\mathcal{N}(0\ ;\ 1)$ si n_1 et n_2 sont supérieurs ou égaux à 30.

s_1 et s_2 sont inconnus mais différents : $T = \dfrac{X_1 - X_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$

f) - la loi de T est approchée par la loi normale $\mathcal{N}(0\ ;\ 1)$ si n_1 et n_2 sont supérieurs ou égaux à 30.

En général, l'hypothèse alternative s'exprime sous l'une des formes suivantes : "$m_1 \neq m_2$" ou "$m_1 < m_2$" ou "$m_1 > m_2$".

Situation 5

Au cours des deux dernières années, une maladie a été traitée dans un hôpital par l'un des deux traitements A ou B. Le choix du traitement pour chaque patient a été fait au hasard.

Les médecins ont observé les résultats suivants :

Protocole thérapeutique	A	B
Nombres de malades traités	300	100
Durée moyenne d'hospitalisation (en jours)	50	53
Variance des durées d'hospitalisation	90	110

i) - Au risque de 5 %, la différence des durées moyennes d'hospitalisation des malades traités par A et B est-elle imputable à une action plus rapide (en moyenne) du traitement A ?

ii) - Au risque de 5 %, la différence des durées moyennes d'hospitalisation des malades traités par A et B est-elle imputable au seul hasard d'échantillonnage ?

Corrigé

On est dans le cas de comparaison de deux moyennes ; l'hypothèse nulle est H_0 : "$m_A = m_B$" où m_A et m_B désignent les durées d'hospitalisation respectives des malades traités par A et B dans la population.

On n'a pas de renseignements sur les variances dans les deux populations, on se place dans le cas **f)**.

Les durées moyennes d'hospitalisation dans les échantillons de patients traités par A et B respectivement peuvent être interprétées comme les réalisations de deux variables aléatoires indépendantes X_A et X_B définies respectivement sur les échantillons de tailles 300 et 100 des deux populations.

La variable de décision est $T = \dfrac{X_A - X_B}{\sqrt{\dfrac{90}{299} + \dfrac{110}{99}}}$ que l'on approche, sous l'hypothèse H_0, par la loi normale $\mathcal{N}(0\,;1)$ car les tailles des deux échantillons sont supérieures à 30. T prend la valeur $T_{obs} = -2{,}52$.

i) - **On teste H_0 : "$m_A = m_B$" contre H_1 : "$m_A < m_B$"** ("les durées sont égales" contre "le traitement A diminue la durée d'hospitalisation")

Le test est unilatéral, le seuil de confiance est $\alpha = 5$ %.

Sous l'hypothèse H_0, le fait que $X_A - X_B$ prenne des valeurs très inférieures à 0 est rare. La valeur critique est le réel $t_{0,05}$ tel que $P[T > t_{0,05}] = 0{,}95$. On trouve $t_{0,05} = -1{,}64$.

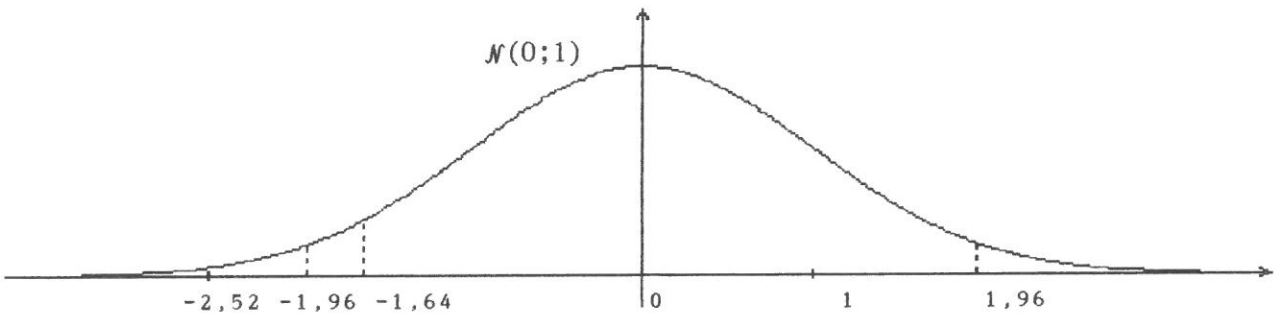


figure 9

D'où $T_{obs} < -1,64$; on n'admet pas l'effet du hasard dans le choix de l'échantillon et on refuse H_0. L'action du traitement A est peut-être en moyenne plus rapide que celle du traitement B, au seuil de 5 %.

ii) - **On teste H_0 : "$m_A = m_B$" contre H_1 : "$m_A \neq m_B$"** ("les durées d'hospitalisation sont égales" contre "les durées sont différentes")

Le test est bilatéral, le seuil de confiance est $\alpha = 5$ %.

Sous l'hypothèse H_0, le fait que $X_A - X_B$ prenne des valeurs éloignées de 0 est rare. La valeur critique est le réel $t'_{0,05}$ tel que :

$$P[|T| < t'_{0,05}] = 0,95.$$

On trouve $t'_{0,05} = 1,96$. D'où $|T_{obs}| > 1,96$; on n'admet pas l'effet du hasard dans le choix de l'échantillon et on refuse H_0. La différence entre les deux moyennes n'est peut-être pas imputable au seul hasard d'échantillonnage, au seuil de 5 %.

Situation 6

Dans une coopérative agricole, on désire tester l'effet d'un engrais sur la production de blé. Pour cela, on choisit 24 lots de terrain de même superficie. La moitié de ces parcelles est traitée avec l'engrais et l'autre moitié ne l'est pas (c'est le groupe de référence). Les autres conditions demeurent les mêmes pour les deux groupes. La moyenne de blé obtenue sur les lots non traités est de 4,8 tonnes avec un écart type de 0,40 tonne, tandis que la moyenne obtenue sur les lots traités est de 5,1 tonnes avec un écart type de 0,36 tonne.

Peut-on conclure qu'il n'y a pas d'amélioration significative de la production de blé avec l'engrais au seuil : **i)** de 1 % **ii)** de 5 %. ?

Corrigé

On est dans le cas de comparaison de deux moyennes ; l'hypothèse nulle est H_0 : "$m_A = m_B$" où m_A et m_B désignent les moyennes des productions de blé des parcelles des types A et B respectivement traitées et non traitées.

On n'a pas de renseignements sur les variances dans les deux populations et les tailles des échantillons sont réduites, on se place dans le cas **c)**.

Les productions moyennes des échantillons A et B peuvent être interprétées comme les réalisations de deux variables aléatoires indépendantes X_A et X_B définies respectivement sur les échantillons de tailles $n_A = n_B = 12$ des deux populations.

La variable de décision est $T = \dfrac{X_A - X_B}{\sigma\sqrt{\dfrac{1}{n_A} + \dfrac{1}{n_B}}}$ où $\sigma^2 = \dfrac{(n_A - 1)\sigma_A^2 + (n_B - 1)\sigma_B^2}{n_A + n_B - 2}$,

$\sigma \approx 0{,}397$; la loi de T est la loi de Student à $n_A + n_B - 2 = 22$ degrés de liberté si le caractère est distribué normalement dans les deux populations, T prend la valeur $T_{obs} = 1{,}85$.

On teste H_0 : "$m_A = m_B$" contre H_1 : "$m_A < m_B$" ("les productions sont égales" contre "l'engrais améliore la production")

Le test est unilatéral, soit α le seuil de signification. Sous l'hypothèse H_0, le fait que $X_A - X_B$ prenne des valeurs "très positives" est rare. La valeur critique est le réel t_α tel que $P[T < t_\alpha] = 1 - \alpha$.

i) - le seuil de confiance est $\alpha = 1$ %.

On trouve $t_{0,01} = 2{,}51$. D'où $T_{obs} < 2{,}51$; on accepte H_0 au seuil de 1 %.

ii) - le seuil de confiance est $\alpha = 5$ %.

On trouve $t_{0,05} = 1{,}72$. D'où $T_{obs} > 1{,}72$; on refuse H_0. L'action de l'engrais augmente peut-être la production au seuil de 5 %.

2°) - Comparaison de deux proportions

Dans deux populations $\mathcal{P}_1$ et $\mathcal{P}_2$, on étudie un caractère qualitatif ayant pour fréquences ou proportions respectives p_1 et p_2 (inconnues).

Il s'agit de savoir, au vu de deux échantillons des deux populations, s'il existe une différence significative entre p_1 et p_2.

La fréquence f_i du caractère dans un échantillon de taille n_i de la population $\mathcal{P}_i$ peut être considérée comme la réalisation d'une variable aléatoires X_i pour $i \in \{1; 2\}$. X_1 et X_2 sont indépendantes.

On élabore un test ; on prend comme hypothèse nulle H_0 : "$p_1 = p_2$".

On pose $f = \dfrac{n_1 f_1 + n_2 f_2}{n_1 + n_2}$ et la variable de décision est $T = \dfrac{X_1 - X_2}{\sqrt{f(1-f)\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$

La loi de T est approchée, sous l'hypothèse H_0, par la loi normale $\mathcal{N}(0\,;\,1)$ lorsque n_1 et n_2 sont supérieurs ou égaux à 30 et que n_1p_1 et n_2p_2 sont supérieurs ou égaux à 15.

En général, l'hypothèse alternative s'exprime sous l'une des formes suivantes : "$p_1 \neq p_2$" ou "$p_1 < p_2$" ou "$p_1 > p_2$".

Situation 7

Pour un sondage électoral, on constitue deux échantillons d'électeurs de tailles 300 et 200 respectivement dans deux circonscriptions A et B. Cela met en évidence des intentions de vote de 56 % et 48 % pour un candidat donné. Tester, au seuil de 5 %, les hypothèses :

i) il y a une différence entre les circonscriptions,

ii) le candidat est préféré dans la circonscription A.

Corrigé

On compare les pourcentages p_A et p_B de votants pour le candidat dans les deux populations A et B respectivement.

Les fréquences 56 % et 48 % de votants pour le candidat dans les deux échantillons peuvent être considérées comme la réalisation de deux variables aléatoires indépendantes X_A et X_B sur les échantillons de tailles 300 et 200 respectivement dans les populations A et B.

L'hypothèse nulle est H_0 : "$p_A = p_B$".

On pose $f = \dfrac{300 \times 0,56 + 200 \times 0,48}{300 + 200} = 0,528$. La variable de décision est

$T = \dfrac{X_A - X_B}{\sqrt{f(1-f)\left(\dfrac{1}{300} + \dfrac{1}{200}\right)}}$ dont on approche la loi, sous l'hypothèse H_0, par $\mathcal{N}(0\,;\,1)$ car 300 et 200 sont supérieurs à 30. T prend la valeur $T_{obs} = 1,75$.

i) - On teste H_0 : "$p_A = p_B$" contre H_1 : "$p_A \neq p_B$"

Le test est bilatéral, le seuil de confiance est $\alpha = 5$ %.

Sous l'hypothèse H_0, le fait que $X_A - X_B$ prenne des valeurs "éloignées" de 0 est rare. La valeur critique est le réel $t_{0,05}$ tel que :

$$P[|T| < t_{0,05}] = 0,95.$$

On trouve $t_{0,05} = 1,96$.

D'où $|T_{obs}| > 1,96$; on n'admet pas l'effet du hasard dans le choix de l'échantillon et on refuse H_0. Il n'y a pas de différence significative entre les deux circonscriptions, au seuil de signification de 5 %.

ii) - On teste H_0 : "$p_A = p_B$" contre H_1 : "$p_A > p_B$"

Le test est unilatéral, le seuil de confiance est $\alpha = 5$ %.

Sous l'hypothèse H_0, le fait que $X_A - X_B$ prenne des valeurs "très supérieures" à 0 est rare. La valeur critique est le réel $t'_{0,05}$ tel que $P[T < t'_{0,05}] = 0,95$.

On trouve $t'_{0,05} = 1,64$.

D'où $T_{obs} > 1,64$; on n'admet pas l'effet du hasard dans le choix de l'échantillon et on refuse H_0. Le candidat n'est sans doute pas préféré, au seuil de 5 %.

De l'observation au modèle théorique

Le test du Khi-2

A - Le problème

On cherche à savoir, à l'aide d'un échantillon, si la distribution d'un caractère dans une population suit une loi donnée, cette loi étant dictée soit par des connaissances sur la population, soit par l'observation de l'échantillon.

B - Quelques exemples

Situation 1 (lois de Mendel)

Un croisement entre roses rouges et blanches a donné en seconde génération des roses rouges, roses et blanches.

Sur un échantillon de taille 600, on a trouvé les résultats suivants :

couleur	effectif
rouge	141
rose	315
blanche	144

Echantillon observé

Les lois de Mendel donnent, en seconde génération : 25 % de roses rouges, 50 % de roses roses et 25 % de roses blanches.

Sur un échantillon de taille 600, on obtient :

couleur	effectif
rouge	150
rose	300
blanche	150

Modèle théorique

On a donc le tableau suivant :

X_i (observé)	x_i (théorique)
141	150
315	300
144	150

Situation 2 (loi de Poisson)

La distribution statistique ci-après donne la répartition de jours sans accidents, avec un accident, ..., avec quatre accidents pour une période de $n = 50$ jours dans une certaine ville.

nombre d'accidents	nombre de jours
0	21
1	18
2	7
3	3
4	1

Echantillon observé

On trouve : $E(X) = 0,9$ et $V(X) = 0,97$ d'où $E(X) \approx V(X)$. On essaie donc d'ajuster la distribution du nombre d'accidents par jour par la loi de Poisson de paramètre 0,9, soit $\mathcal{P}(0,9)$.

En utilisant la table de la loi de Poisson, on a pour $\mathcal{P}(0,9)$ et pour un échantillon d'effectif 50 :

nombre d'accidents	probabilités	nombre de jours
0	0,4066	20,330
1	0,3659	18,295
2	0,1647	8,235
3	0,0494	2,470
4	0,0111	0,555

Modèle théorique

On a donc le tableau suivant :

X_i (observé)	x_i (théorique)
21	20,330
18	18,295
7	8,235
3	2,470
1	0,555

Situation 3 (loi normale)

On fait passer un examen de mathématiques à un groupe de 300 étudiants. On obtient le tableau de notes suivant :

note x	nombre d'étudiants
$0 \leqslant x < 4$	8
$4 \leqslant x < 8$	52
$8 \leqslant x < 12$	102
$12 \leqslant x < 16$	96
$16 \leqslant x \leqslant 20$	42

Echantillon observé

La question est de savoir si la distribution des étudiants peut-être considérée comme normale.

On calcule la moyenne et l'écart type de cette série statistique, on trouve : $m = 11,5$ et $\sigma = 4$. Sous l'hypothèse d'une distribution normale, la variable

aléatoire égale à la note obtenue suit une loi normale $\mathcal{N}(11,5\ ;\ 4)$ et la variable aléatoire $T = \frac{X - 11,5}{4}$ suit la loi normale $\mathcal{N}(0\ ;\ 1)$

En utilisant la table de la fonction de répartition de la loi normale centrée réduite, on obtient le tableau suivant :

classes	probabilités	effectifs théoriques
$x < 0$	0,002	0,6
$0 \leqslant x < 4$	0,028	8,4
$4 \leqslant x < 8$	0,159	47,7
$8 \leqslant x < 12$	0,359	107,7
$12 \leqslant x < 16$	0,321	96,3
$16 \leqslant x < 20$	0,118	35,4
$x \geqslant 20$	0,013	3,9

Modèle théorique

On regroupe les deux premières et les deux dernières classes.

On a donc le tableau suivant :

X_i (observé)	x_i (théorique)
8	9
32	47,7
102	107,7
96	96,3
42	39,3

B - Mise en place du test du χ^2

1°) - Le côté mathématique

Dans un échantillon de taille n, le caractère étudié - qualitatif ou quantitatif - se répartit en N classes d'effectif X_i auquel correspond un effectif théorique x_i donné par la distribution que l'on cherche à tester.

L'écart entre la répartition observée et la répartition théorique est mesuré par l'écart quadratique réduit $\sum_{i=1}^{N} \frac{(X_i - x_i)^2}{x_i}$; c'est χ^2_{obs}, le <u>khi-deux observé</u> ou <u>indicateur d'écart</u>. Les X_i peuvent être considérées comme des variables aléatoires liées par la relation $X_1 + X_2 + \ldots + X_N = n$.

$\chi^2 = \sum_{i=1}^{N} \frac{(X_i - x_i)^2}{x_i}$ est une variable aléatoire qui a une distribution en khi-carré à ν degrés de liberté où :

* $\nu = N - 1$ si les effectifs théoriques peuvent être calculés sans qu'aucun paramètre de la population n'est à être estimé à l'aide des statistiques sur échantillon,
* $\nu = N - 1 - k$ si les effectifs espérés ne peuvent être calculés qu'à l'aide de l'estimation de k paramètres de la population.

La densité de la loi du khi-carré à ν degrés de liberté est f définie par :

$$f(x) = \frac{x^{\left(\frac{\nu}{2}-1\right)}}{2^{\frac{\nu}{2}}\,\Gamma\left(\frac{\nu}{2}\right)} e^{-\frac{x}{2}} \mathbb{1}_{\mathbb{R}^+}(x)$$ où Γ désigne la fonction "Gamma",

$\Gamma(a) = \int_0^{\infty} e^{-t}\, t^{a-1}\, dt$, en particulier pour $n \in \mathbb{N}^*$, $\Gamma(n) = (n-1)!$.

On fixe α le <u>seuil de confiance</u> du test, c'est-à-dire le risque de rejeter l'hypothèse d'ajustement alors qu'elle est vraie.
On détermine la valeur critique χ^2_α telle que $P[\chi^2 > t_\alpha] = \alpha$.
Si on suppose que la distribution de la population suit la loi théorique, le fait d'avoir un khi-deux important est rare, ce qui conduit au raisonnement suivant :

- si $\chi^2_{obs} \geqslant \chi^2_\alpha$: ce fait étant relativement peu probable sous l'hypothèse d'ajustement, on n'admet pas l'effet du hasard dans le choix de l'échantillon et on rejette l'hypothèse d'ajustement,
- si $\chi^2_{obs} < \chi^2_\alpha$: on accepte l'hypothèse d'ajustement ; au seuil de signification α, la différence entre les distributions observée et théorique n'est pas significative.

2°) - Réalisation du test

a) - Calcul de χ^2_{obs}

Pour que cet indicateur ait un sens, il est nécessaire que x_i ne soit pas trop petit. On admet que x_i doit être supérieur à 5 ; dans le cas contraire, on regroupe plusieurs classes adjacentes de façon à obtenir $x_i \geqslant 5$.
Soit alors N le nombre de valeurs distinctes de X_i (et donc de x_i) ; par définition , $\chi^2_{obs} = \sum_{i=1}^{N} \frac{(X_i - x_i)^2}{x_i}$.

Situation 1 : On obtient $N = 3$; $\chi^2_{obs} = 1,53$.

Situation 2 : On regroupe les trois dernières classes, on obtient :

X_i (observé)	x_i (théorique)
21	20,330
18	18,295
11	11,260

$N = 3$; $\chi^2_{obs} = 0,0328$.

Situation 3 : On obtient $N = 5$; $\chi^2_{obs} = 0,987$.

b) - **Calcul du nombre de degrés de liberté**

Situation 1 : On ne fait pas intervenir de paramètres ; $\nu = N - 1 = 2$.

Situation 2 : On utilise la loi de Poisson de paramètre 0,9 qui fait donc intervenir 1 paramètre estimé de la population (la moyenne) ; $\nu = N - 1 - 1 = 1$.

Situation 3 : On utilise la loi normale $\mathcal{N}(11,5 ; 4)$ qui fait intervenir 2 paramètres estimés de la population (moyenne et écart type) ; $\nu = N - 1 - 2 = 2$.

c) - On fixe un seuil de signification α (en général 5 % ou 1 %).

d) - A l'aide de tables, on calcule χ^2_α, la valeur critique, qui dépend de α et du nombre ν de degrés de liberté.

e) - On applique la règle de décision.

Situation 1 : $\alpha = 5$ % ; $\nu = 2$; $\chi^2_{0,05} = 5,99$ (table) comme $\chi^2_{obs} = 1,53$, on accepte l'ajustement.

Situation 2 : $\alpha = 5$ % ; $\nu = 1$; $\chi^2_{0,05} = 3,94$ (table) comme $\chi^2_{obs} = 0,0328$, on accepte l'ajustement.

Situation 3 : $\alpha = 5$ % ; $\nu = 2$; $\chi^2_{0,05} = 5,99$ (table) comme $\chi^2_{obs} = 0,987$, on accepte l'ajustement.

3°) - **Des exercices**

Situation 4 :

Un couple de cobayes à pelage gris et lisse a donné naissance à 128 descendants dont les pelages se répartissent de la manière suivante :

78 au pelage gris et lisse (*gl*) ; 19 au pelage blanc et rude (*br*)
26 au pelage blanc et lisse (*bl*) ; 5 au pelage gris et rude (*gr*)

Si l'on admet que la transmission de ces deux caractères suit une loi de Mendel, les fréquences théoriques d'apparition des différentes races doivent être :

$P(gl) = \frac{9}{16}$; $P(bl) = \frac{3}{16}$; $P(br) = \frac{3}{16}$; $P(gr) = \frac{1}{16}$.

Les résultats expérimentaux permettent-ils d'accepter ce mode de transmission ?

Corrigé :

On obtient le tableau :

X_i (observé)	78	26	19	5
x_i (théorique)	72	24	24	8

$\alpha = 5$ % ; $\nu = 4 - 1 = 3$; $\chi^2_{0,05} = 7,81$ (table) comme $\chi^2_{obs} = 2,833$, on accepte l'ajustement.

Situation 5 :

On a enregistré pendant $n = 1\ 000$ jours le nombre d'appels reçus par un vétérinaire de garde entre 20 h et 6 h du matin. On a obtenu les résultats suivants :

Nombre d'appels	0	1	2	3	4	5	6	7	8	9	10	11
Nombre de jours	14	70	155	185	205	150	115	65	30	5	1	5

1°) - Déterminer la moyenne et l'écart type de cette série statistique.

2°) - Tester à l'aide du χ^2, au risque de 5 %, l'hypothèse : "Le nombre d'appels reçus par le vétérinaire de garde suit une distribution de Poisson".

Corrigé :

$E(X) = 4$, $V(X) = 3,74$. On peut supposer que $X \sim \mathcal{P}(4)$.

On obtient le tableau :

X_i (observé)	14	70	155	185	205	150	115	65	30	5	1	5
x_i (théorique)	18	73	146	195	195	156	104	60	30	13	5	2

On doit regrouper les deux dernières classes.

$\alpha = 5$ % ; $\nu = 11 - 1 - 1 = 9$; $\chi^2_{0,05} = 16,92$ (table) comme $\chi^2_{obs} = 9,47$, on accepte l'ajustement.

Situation 6 :

On étudie la taille de 300 individus tirés au hasard dans une population de très grand effectif. On obtient les résultats suivants :

Taille	160	162	164	166	168	170	172	174	176	178	180	182
Effectifs	4	1	25	35	75	115	125	60	40	10	5	5

Sachant que les classes de tailles sont représentées dans le tableau ci-dessus par leur valeur centrale, tester, au risque de 5 %, l'hypothèse : "La taille des 500 individus suit une loi normale".

Corrigé :

$E(X) = 170,86$, $V(X) = 3,6$. On peut supooser que $X \sim \mathcal{N}(170,86 ; 3,6)$.

On obtient le tableau :

X_i (observé)	5	25	35	75	115	125	60	40	10
x_i (théorique)	7,30	18,50	45,35	79,60	107,25	103,20	76,25	40,70	15,85

On a regroupé les classes extrêmes par deux.

$\alpha = 5$ % ; $\nu = 10 - 1 - 2 = 7$; $\chi^2_{0,05} = 14,07$ (table) comme $\chi^2_{obs} = 19,19$, on rejette l'ajustement.

Ajustement graphique par une loi normale : la droite de Henry

I. Rappels de statistique descriptive

La *fonction de répartition* d'une série statistique continue est *l'application,* qui associe à tout nombre réel x, la proportion des individus de la population pour lesquels la valeur du caractère observé est inférieure ou égale à x. Usuellement, cette fonction est notée F, c'est une fonction croissante, affine par intervalles. En effet, si la variable statistique continue x prend toute valeur d'un intervalle de nombres réels $[e_0, e_n[$ partagé en n classes du type $[e_i, e_{i+1}[$, pour i entier naturel variant de 0 à n - 1, on a : si $x < e_0$: $F(x) = 0$,

si $e_i \leq x < e_{i+1}$: $F(x) = \sum_{k=1}^{i} c_i f_k + \frac{x - e_i}{e_{i+1} - e_i} \cdot f_{i+1}$ pour i variant de 0 à n - 1, où f_{i+1} est la fréquence de la classe $[e_i, e_{i+1}[$,

si $x \geq e_n$: $F(x) = \sum_{k=1}^{n} c_i f_k = 1$.

La représentation graphique de la fonction de répartition est appelée ***courbe cumulative*** de la série statistique.

Exemple :

Dans une banque, on a relevé les montants des retraits en espèces effectués par mille clients sur une période de un mois. On a obtenu les résultats suivants :

Montant des retraits exprimés en francs	Nombre de clients	Fréquence	Fréquences cumulées croissantes
moins de 500	5	0,005	0,005
de 500 à moins de 1000	12	0,012	0,017
de 1000 à moins de 1500	33	0,033	0,050
de 1500 à moins de 2000	71	0,071	0,121
de 2000 à moins de 2500	119	0,119	0,240
de 2500 à moins de 3000	175	0,175	0,415
de 3000 à moins de 3500	185	0,185	0,600
de 3500 à moins de 4000	158	0,158	0,758
de 4000 à moins de 4500	122	0,122	0,880
de 4500 à moins de 5000	69	0,069	0,949
de 5000 à moins de 5500	35	0,035	0,984
de 5500 à moins de 6000	11	0,011	0,995
6000 et plus	5	0,005	1
TOTAL	**1000**	**1**	

(source confidentielle)

Dans cet exemple la population observée est l'ensemble des mille clients suivant le caractère "montant des retraits effectués en espèces au cours du mois considéré". Il s'agit d'un caractère quantitatif continu selon la convention : bien que les valeurs observées soient des nombres entiers

ou des nombres décimaux d'ordre deux, la variable statistique associée sera traitée comme une variable continue pouvant donc prendre toute valeur d'un intervalle de nombres réels.

La variable statistique associée fait correspondre à chaque client la modalité retenue suivant le montant de ses retraits.

La série statistique présentée associe à chaque modalité [0,500[, [500,1000[, ..., [6000,+∞[le nombre des clients dont les retraits au cours du mois sont situés dans cet intervalle.

***Histogramme des fréquences* :**

Dans un repère orthogonal, les classes qui constituent les modalités de la série statistique observée sont représentées par des segments portés par l'axe des abscisses, sur lesquels sont construits des rectangles *d'aires proportionnelles aux fréquences* des classes correspondantes.

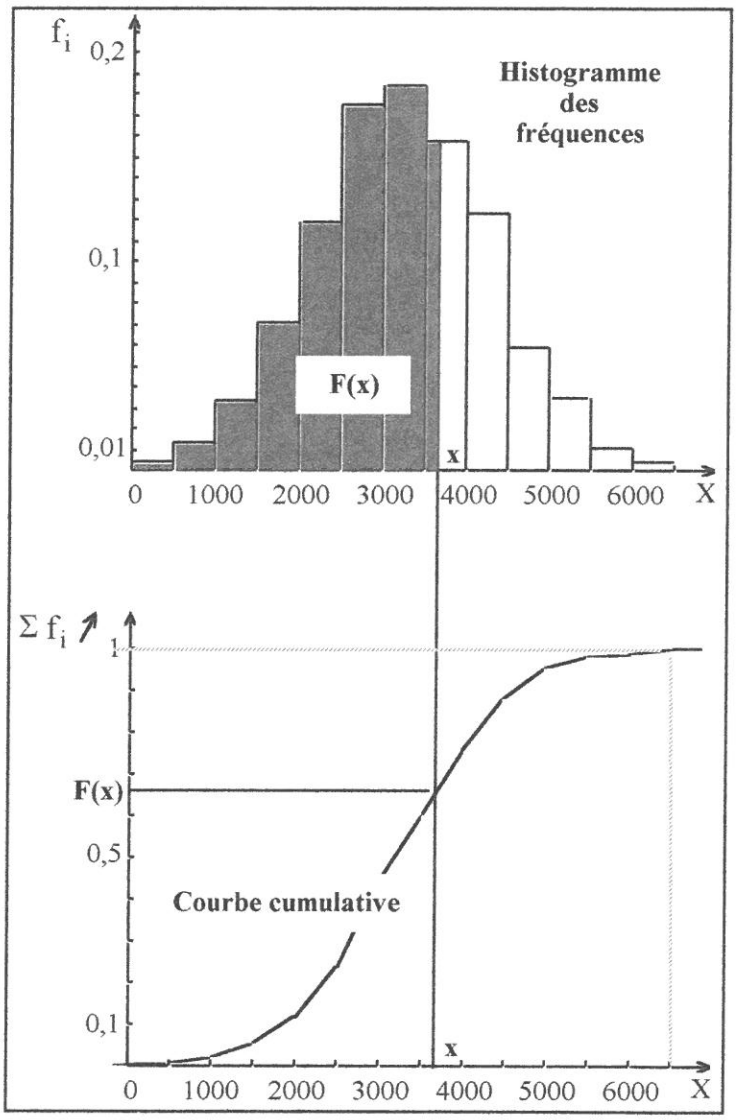
Dans cet exemple, les douze premières classes ont la même amplitude, la dernière classe [6000,+∞[est alors représentée par convention, avec cette même amplitude.

Représentation graphique de la fonction de répartition de cette série statistique et lien avec l'histogramme des fréquences de cette série statistique continue : l'aire de la partie grisée de l'histogramme des fréquences est égale à F(x).

La fonction F est dérivable par intervalles, sa dérivée est la fonction en escalier f telle que, pour tout x de $[e_i, e_{i+1}[$:

$$f(x) = \frac{f_{i+1}}{e_{i+1} - e_i}.$$

Si l'unité graphique choisie sur l'axe des abscisses est telle que $e_{i+1} - e_i = 1$, alors la représentation graphique de f est la frontière supérieure de l'histogramme des fréquences, que l'on appelle parfois pour cela "diagramme différentiel" de la série statistique, la courbe cumulative correspondante est alors qualifiée de "diagramme intégral".



II. Droite de Henry : le principe.

Méthode graphique permettant de tester si une distribution statistique, qui engendre une variable aléatoire X lorsque les fréquences observées sont assimilées à des probabilités, est susceptible d'être ajustée par une loi N(m, σ).

- **Fonction de répartition de la série statistique:**

  Soit une série statistique à caractère quantitatif qui permet de définir une variable aléatoire X et F la fonction de répartition de la variable statistique. Pour tout réel x, F(x) est la fréquence cumulée croissante jusque x.

- **Fonction de répartition de la loi normale :**

  X suit la loi normale de paramètres m et s si et seulement si $T = \frac{X - m}{s}$ suit la loi N(0,1).

  Pour tout réel x, soit p la probabilité de l'événement $P(X \leq x)$. On a alors :

$$p = P(X \leq x) = P\left(T \leq \frac{x - m}{\sigma}\right) = \pi\left(\frac{x - m}{\sigma}\right)$$

  La fonction π est une bijection de **R** sur]0,1[, d'où : $\frac{x - m}{\sigma} = \pi^{-1}(p)$ soit $\pi^{-1}(p) = \frac{1}{\sigma}x - \frac{m}{\sigma}$.

  On en déduit que dans un repère donné, l'ensemble des points de coordonnées $(x, \pi^{-1}(p))$ est la droite d'équation $y = \frac{1}{\sigma}x - \frac{m}{\sigma}$.

- **Comparaison des deux répartitions :**

  Les fonctions de répartition de la variable statistique et de la variable aléatoire X sont égales si et seulement si pour tout x de **R** : $F(x) = P(X \leq x) = \pi\left(\frac{x - m}{\sigma}\right) = \pi(y)$ avec $y = \frac{x - m}{\sigma}$.

  Vérifier la normalité de la série statistique est donc équivalent à montrer l'existence de deux nombres m et σ, tels que $F(x) = \pi(y)$ avec $y = \frac{1}{\sigma}x - \frac{m}{\sigma}$.

On observe graphiquement que la série statistique peut être ajustée par une loi normale de paramètres m et σ si les points de coordonnées $(x, \pi^{-1}(F(x)))$ sont sensiblement alignés. On précise l'alignement au besoin par la méthode des moindres carrés.

III. Construction et utilisation du papier gausso-arithmétique.

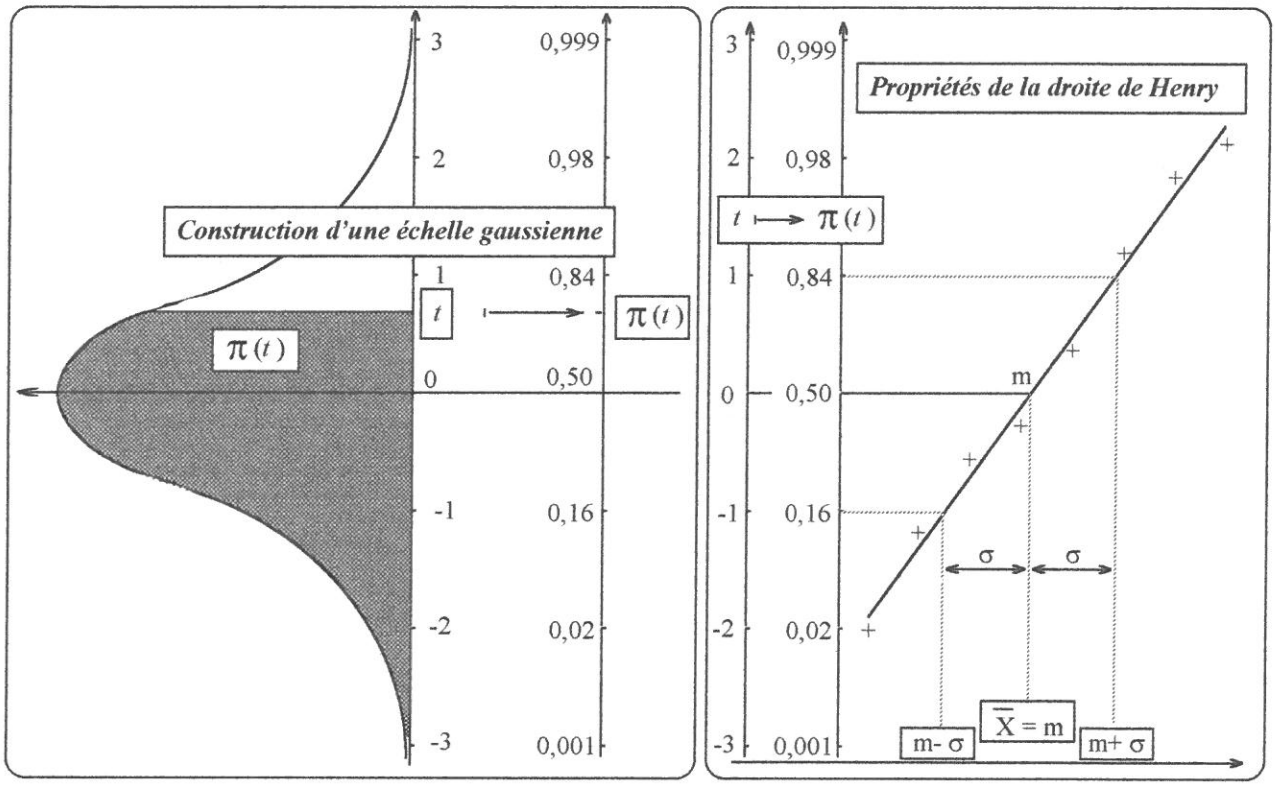
On porte en abscisse les valeurs e_{i+1} des limites supérieures des classes de la série statistique et en ordonnée $\pi^{-1}(F(e_{i+1}))$.

La loi est gaussienne si et seulement si le nuage des points de coordonnées $M(e_{i+1}, \pi^{-1}(F(e_{i+1})))$ est sensiblement rectiligne.

L'utilisation du papier gausso-arithmétique évite le calcul des nombres $\pi^{-1}(F(e_{i+1}))$. La graduation des abscisses est arithmétique, celle des ordonnées est fonctionnelle, dite gaussienne.

La droite d'ajustement affine, tracée "au jugé" est appelée droite de HENRY.

Le papier gausso-arithmétique est fabriqué par la Compagnie Française des Diagrammes (22, Boulevard d'Inkermann 92200 NEUILLY-SUR-SEINE). On en trouve dans les librairies spécialisées.



IV Propriétés graphiques immédiates.

On exploite la relation $F(x) = \pi(y)$ avec $y = \frac{1}{\sigma}x - \frac{m}{\sigma}$.

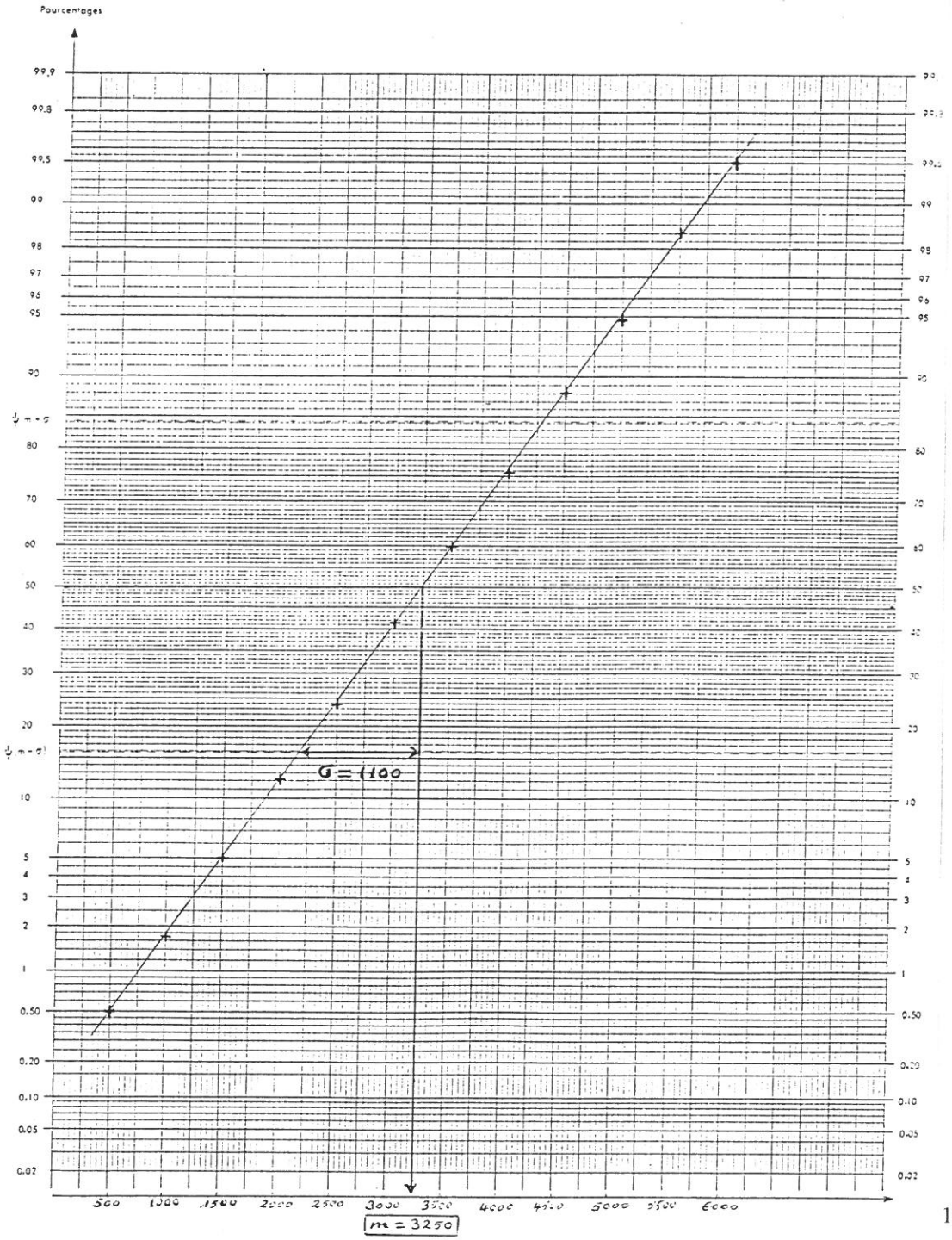
- Puisque $\pi^{-1}(0,5) = 0$ et que $y = 0,5$ lorsque $x = m$, le point d'ordonnée 0,5 de la droite de Henry a pour abscisse la moyenne m de la variable aléatoire X.
- Puisque $\pi^{-1}(1) \approx 0,84134$ et que $y = 1$ lorsque $x = m + \sigma$, le point d'ordonnée 0,84 de la droite de Henry a pour abscisse $m + \sigma$.
- Puisque $\pi^{-1}(-1) \approx 0,158664$ et que $y = -1$ lorsque $x = m - \sigma$, le point d'ordonnée 0,16 de la droite de Henry a pour abscisse $m - \sigma$.

Ces observations permettent d'obtenir rapidement, par lecture graphique, une approximation de la moyenne et de l'écart type de la variable aléatoire X réalisant une approximation de la variable statistique étudiée.

Exemple :

Observons les montants des retraits en espèces effectués par mille clients d'une banque, statistique dont le tableau des fréquences cumulées croissantes est donné en introduction.

La droite de Henry tracée "au jugé" sur le graphique suivant permet d'affirmer que la variable aléatoire qui associe à chaque client le montant de ses retraits mensuels peut être approchée par une loi normale, de paramètres $m = 3\,250$ et $\sigma = 1\,100$.



V. Une activité aléatoire à propos de la fonction *randomise* des calculatrices.

La fonction randomise, notée "RAN #" sur les calculatrices, génère "au hasard" des nombres de l'intervalle [0,1]. Cette fonction est une variable aléatoire qui suit, théoriquement, la loi uniforme sur l'ensemble des nombres décimaux d'ordre 10 de l'intervalle [0, 1], chacun de ces nombres ayant la même probabilité d'être obtenu. Ces $(10^{10} + 1)$ nombres constituent une population de moyenne m et d'écart type σ dont le calcul nécessite de connaître au préalable les deux égalités :

$$\sum_{k=1}^{n} k = \frac{n(n+1)}{2} \quad \text{et} \quad \sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}$$

$$m = \frac{1}{10^{10}+1} \sum_{k=0}^{10^{10}} k\, 10^{-10} = \frac{10^{-10}}{10^{10}+1} \sum_{k=0}^{10^{10}} k = \frac{10^{-10}}{10^{10}+1} \cdot \frac{10^{10}(10^{10}+1)}{2} = \frac{1}{2}.$$

$$\sigma^2 = \frac{1}{10^{10}+1} \sum_{k=0}^{10^{10}} (k\, 10^{-10})^2 - m^2 = \frac{10^{-20}}{10^{10}+1} \sum_{k=0}^{10^{10}} k^2 - \left(\frac{1}{2}\right)^2 = \frac{10^{-20}}{10^{10}+1} \cdot \frac{10^{10}(10^{10}+1)(2.10^{10}+1)}{6} - \frac{1}{4}$$

$$= \frac{10^{-10}(2.10^{10}+1)}{6} - \frac{1}{4} = \frac{1}{3} - \frac{1}{4} - \frac{10^{-10}}{6} \approx \frac{1}{12}. \qquad \text{D'où} \quad \sigma \approx 0{,}29.$$

Remarque **:** Les nombres décimaux d'ordre 10 de l'intervalle [0, 1] sont suffisamment nombreux pour que l'on puisse approcher la loi de la variable aléatoire discrète RAN # par la loi uniforme continue sur [0, 1]. Les calculs sont alors simplifiés, on obtient :

$$m = \int_0^1 x\, dx = \left[\frac{x^2}{2}\right]_0^1 = \frac{1}{2} \quad \text{et} \quad \sigma^2 = \int_0^1 x^2\, dx - m^2 = \left[\frac{x^3}{3}\right]_0^1 - \frac{1}{4} = \frac{1}{12}.$$

On dispose donc d'une population de moyenne $m = 0{,}5$ et d'écart type $\sigma \approx 0{,}29$.

Une conséquence du théorème de la limite centrée est : pour n suffisamment grand, la variable aléatoire $\overline{X}$ qui, à tout échantillon non exhaustif de taille n prélevé dans cette population, associe la moyenne de cet échantillon, suit approximativement la loi normale $N\left(m, \frac{\sigma}{\sqrt{n}}\right)$, ce que l'on se propose de vérifier concrètement, par exemple avec $n = 30$.

Théoriquement, $\overline{X}$ suit alors approximativement la loi normale N(0,5 ; 0,053).

Le programme de calcul suivant permet d'obtenir 100 valeurs prises par la variable aléatoire $\overline{X}$ rangées en ordre croissant pour une exploitation commode.

```
10   VAC : DIM X(100)
20   FOR I = 1 TO 100
30   FOR J = 1 TO 30
40   Z = Z + RAN #
50   NEXT J
60   X(I) = Z/30 : Z = 0
70   NEXT I
80   FOR I = 1 TO 100 N = 100 - I
90   FOR K = 1 TO N
100  IF X(I) > X(I+K)
     THEN U = X(I) : X(I) = X(I+K) : X(I+K) = U
110  NEXT K
120  NEXT I
130  FOR I = 1 TO 100
140  PRINT X(I)
150  NEXT
```

Résultats obtenus :

0,3674...	0,3808...	0,3811...	0,3882...	0,3991...	0,4041...	0,4065...	0,4088...	0,4094...	0,4203...
0,4228...	0,4253...	0,4268...	0,4360...	0,4363...	0,4377...	0,4419...	0,4420...	0,4441...	0,4458...
0,4480...	0,4490...	0,4499...	0,4505...	0,4530...	0,4568...	0,4621...	0,4624...	0,4640...	0,4660...
0,4666...	0,4708...	0,4711...	0,4723...	0,4729...	0,4758...	0,4762...	0,4782...	0,4801...	0,4831...
0,4835...	0,4862...	0,4865....	0,4870...	0,4872...	0,4882...	0,4901...	0,4906...	0,4908...	0,4914...
0,4925...	0,4929...	0,4960...	0,4975...	0,4987...	0,4995...	0,5030...	0,5037...	0,5059...	0,5064...
0,5077...	0,5092...	0,5104...	0,5111...	0,5122...	0,5153...	0,5159...	0,5169...	0,5174...	0,5181...
0,5190...	0,5198...	0,5211...	0,5219...	0,5238...	0,5262...	0,5263...	0,5282...	0,5288...	0,5348...
0,5362...	0,5386...	0,5413...	0,5425...	0,5454...	0,5472...	0,5478...	0,5542...	0,5560...	0,5591...
0,5599...	0,5644...	0,5647...	0,5668...	0,5680...	0,5704...	0,5811...	0,5877...	0,6314...	0,6636...

Regroupement des données en classes :

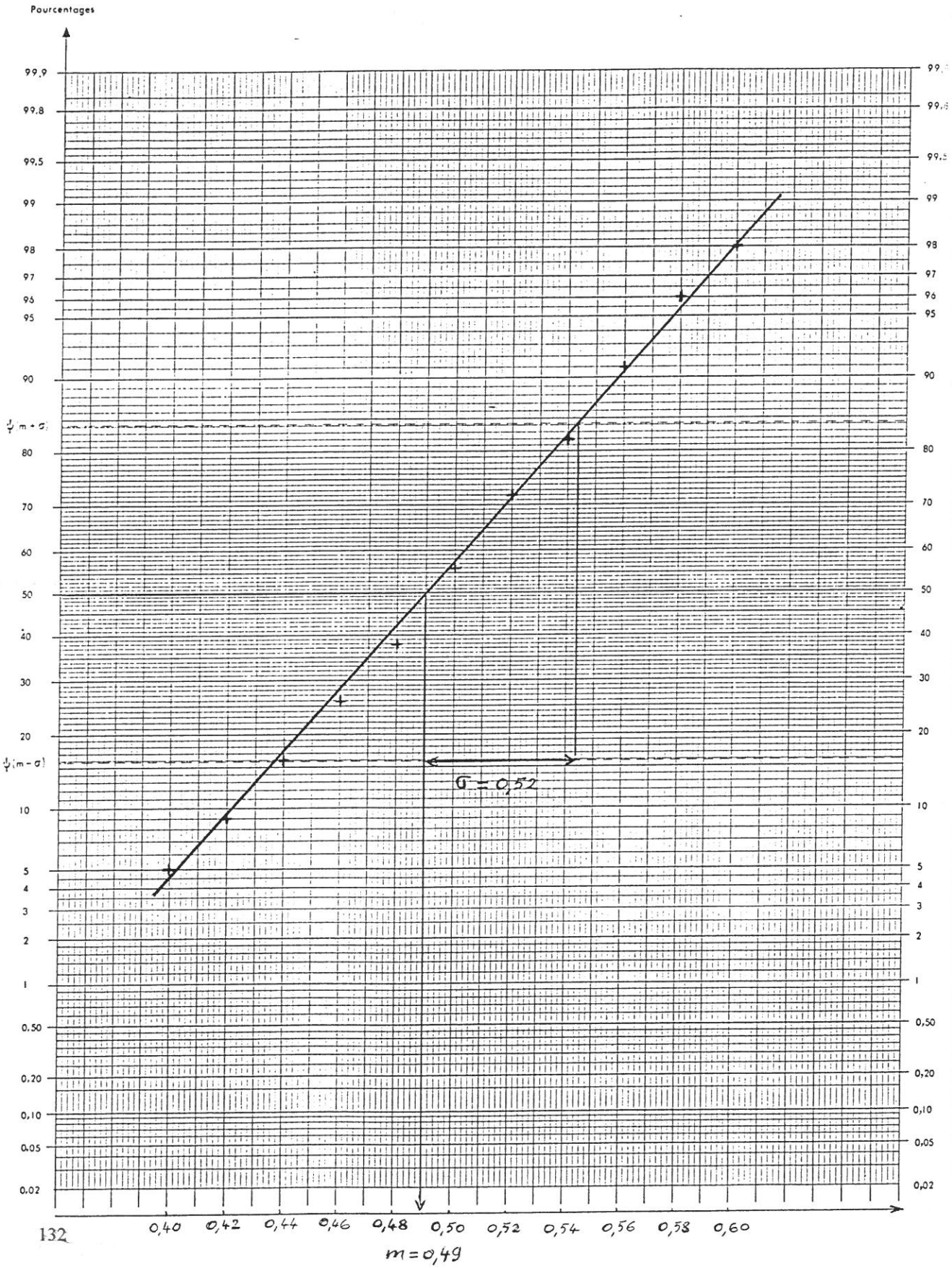
Classes	Effectifs	Fréquences cumulées croissantes
] -∞ ; 0,40 [	5	0,05
[0,40 ; 0,42 [	4	0,09
[0,42 ; 0,44 [	7	0,16
[0,44 ; 0,46 [	10	0,26
[0,46 ; 0,48 [	12	0,38
[0,48 ; 0,50 [	18	0,56
[0,50 ; 0,52 [	16	0,72
[0,52 ; 0,54 [	10	0,82
[0,54 ; 0,56 [	9	0,91
[0,56 ; 0,58 [	5	0,96
[0,58 ; 0,60 [	2	0,98
[0,60 ; +∞ [	2	1

On observe sur le graphique suivant que le nuage des points de coordonnées $(e_{i+1}, \pi^{-1}(F(e_{i+1})))$ est susceptible d'un ajustement affine, ce qui justifie que l'on puisse approcher la loi de $\bar{X}$ par une loi normale.

La droite de Henry ajustant ce nuage de points permet de constater alors par lecture graphique, que la variable aléatoire $\bar{X}$ qui, à tout échantillon non exhaustif de taille 30 prélevé dans cette population, associe la moyenne de cet échantillon, suit approximativement la loi normale N(0,49, 0,052), ce qui confirme "expérimentalement" le résultat obtenu par le théorème de la limite centrée.

On observe sur le graphique suivant que le nuage des points $M(e_{i+1}, \pi^{-1}(F(e_{i+1})))$ est susceptible d'un ajustement affine, ce qui justifie que l'on puisse approcher la loi de $\overline{X}$ par une loi normale.

La droite de Henry ajustant ce nuage de points permet de constater alors par lecture graphique, que la variable aléatoire $\overline{X}$ qui, à tout échantillon non exhaustif de taille 30 prélevé dans cette population, associe la moyenne de cet échantillon, suit approximativement la loi normale $\mathcal{N}(0{,}49 \ ; \ 0{,}052)$, ce qui confirme "expérimentalement" le résultat obtenu par le théorème de la limite centrée.



Fiabilité

I. Historique du développement des concepts relatifs à la fiabilité

Depuis l'origine de l'homme

Depuis qu'il a conçu et utilisé ses premiers outils, l'Homme s'est préoccupé de leur fiabilité.

Dans la première moitié du XX^{e} siècle

"Une chaîne est aussi forte que son maillon le plus faible."

Le mathématicien PIERUSKA rectifie cette certitude : "Si la probabilité de survie d'un élément est $\frac{1}{x}$, la probabilité de survie de n éléments identiques en série est $\frac{1}{x^n}$."

La fiabilité d'un élément doit être beaucoup plus élevée que la fiabilité exigée du système.

Dans les années 40

Loi de MURPHY : "If anything can go wrong, it will !" (Si un ennui a la moindre chance de se produire, dites-vous qu'il se produira !)

Pour des raisons évidentes, cette loi est souvent appelée "loi de l'emmerdement maximal" !

A cette époque, le manque de fiabilité est devenu le cauchemar des ingénieurs. Aux Etats-Unis, General Motors Corporation réussit à étendre la durée de vie utile de ses moteurs de traction de locomotives de 250 000 miles à 1 000 000 miles.

Dans les années 50

L'idée naît qu'il est plus raisonnable de concevoir des équipements fiables plutôt que d'attendre la défaillance et de réparer ensuite !

La révolution électronique et la miniaturisation des circuits conduit à la création de l'"Advisory Group on Reliability of Electronic Equipments".

On arrive à des composants de très forte fiabilité. Les normes en sont définies, reprises par la N.A.S.A. et le C.N.E.T.

Dans les années 60

La fiabilité apparaît dans des secteurs de plus en plus variés. Citons deux grands pôles :

- l'aéronautique et l'aérospatiale. En particulier, en France, la SNIAS introduit la Méthode des Combinaisons de Pannes pour ses projets Concorde puis Airbus,
- la construction de centrales nucléaires, où la prise en compte d'accidents potentiels est portée à un niveau sans précédent dans l'industrie.

C'est dans ces années que les aspects mathématiques de la fiabilité se développent, grâce à : BASOVSKY, BARLOW, BIRNAUM, PROSCHAN et WEIBULL.

En France, c'est en 1962 que le mot "fiabilité" est admis à l'Académie des Sciences.

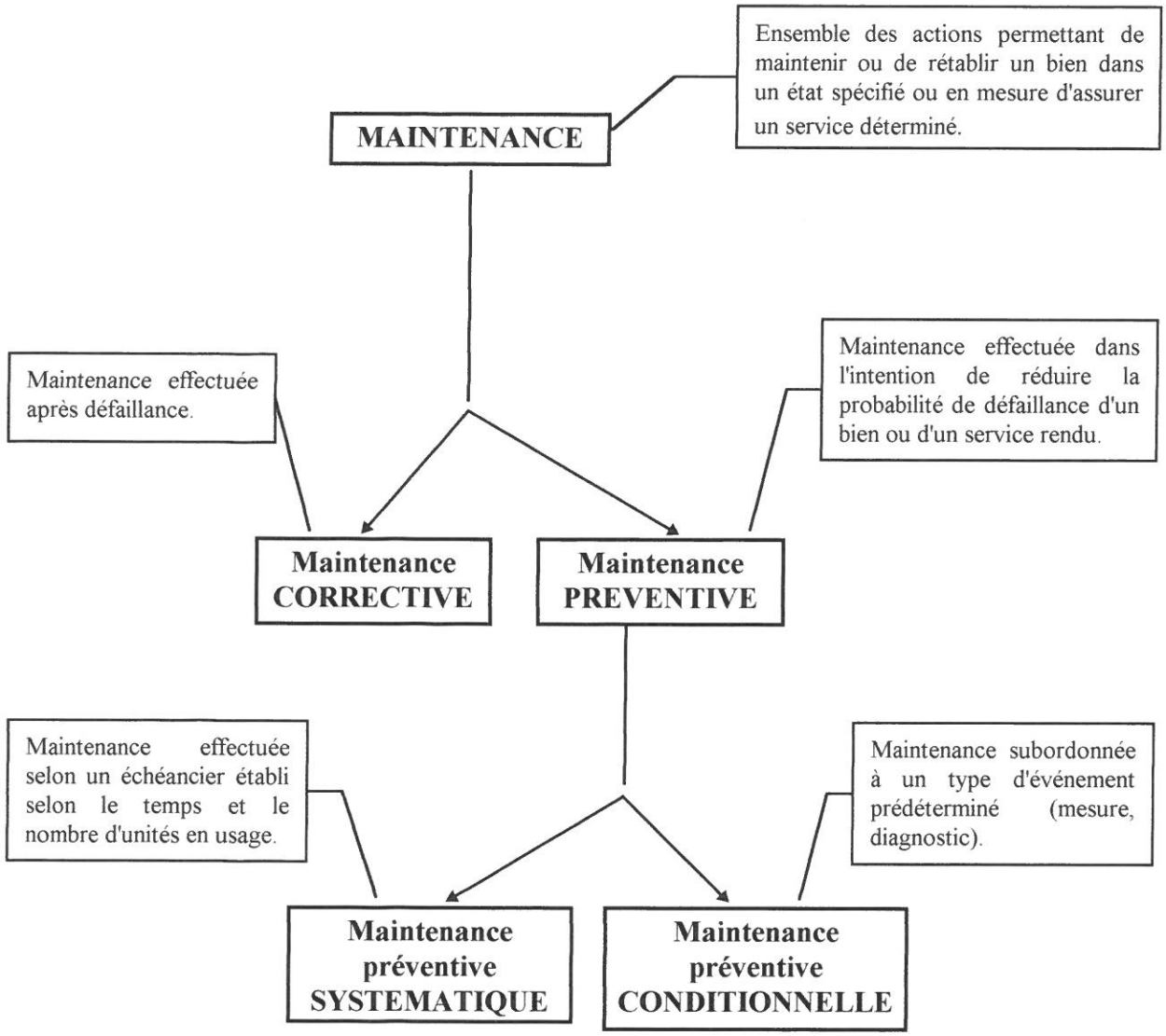
Dans les années 70

On assiste à des développements tous azimuts :

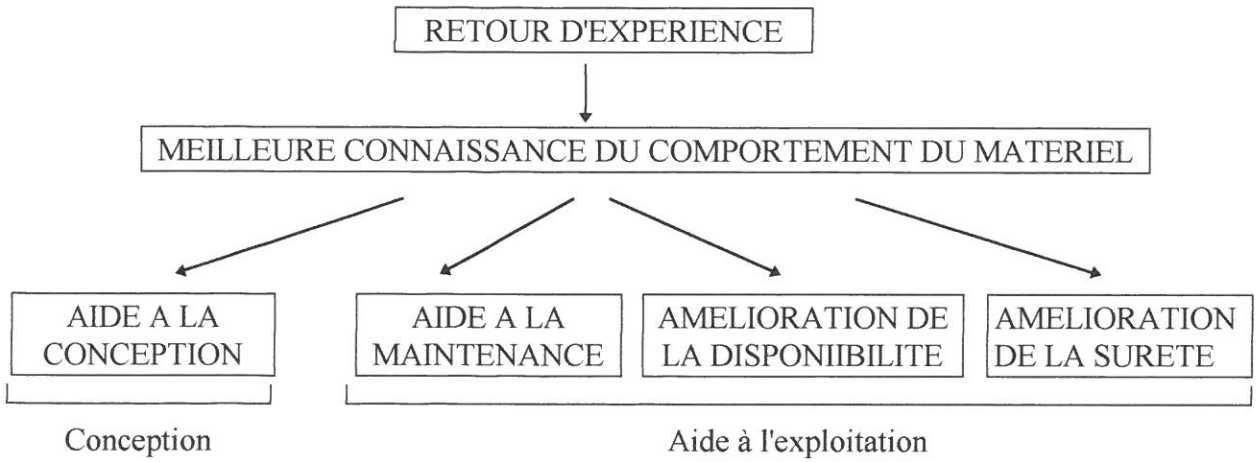
- prévision des risques industriels, (entre autres : caractérisation du risque nucléaire),
- fiabilité des logiciels,
- fiabilité de l'opérateur humain,
- en médecine : * haute fiabilité des installations,
  * méthodes pour l'analyse statistique d'échantillons de durée de vie.

II. Notion de maintenance

La maintenance peut être définie comme l'ensemble des actions permettant de maintenir ou de rétablir un matériel dans un état spécifié ou en mesure d'assurer un service déterminé :



Retour d'expérience



III. La notion de défaillance

La défaillance est la cessation d'un dispositif à accomplir sa (ou ses) fonction(s) requise(s).

On distingue plusieurs types de défaillances :

A_1 : la défaillance progressive

qui aurait dû être prévue par un examen ou une surveillance antérieurs

A_2 : la défaillance soudaine

non prévisible par les examens ou surveillances antérieurs

B_1 : la défaillance ponctuelle

résulte de déviations d'une ou plusieurs caractéristiques du système, mais n'entraîne pas la disparition complète de la fonction requise

B_2 : la défaillance totale

la fonction du dispositif est complètement perdue

$A_1 + B_1$: la défaillance par dégradation

$A_2 + B_2$: la défaillance catalectique

Une défaillance peut être grave pour :

- **la sûreté** : (exemple : centrale nucléaire),
- **la production** : arrêt ou diminution de la production,
- **la maintenance** : dommages conduisant à une réparation coûteuse.

IV. Le modèle mathématique

La fiabilité est la caractéristique d'un dispositif qui s'exprime par la probabilité pour ce dispositif d'accomplir une fonction requise, dans des conditions données, pendant une période donnée (définition AFNOR).

A. Fonction de fiabilité, fonction de défaillance

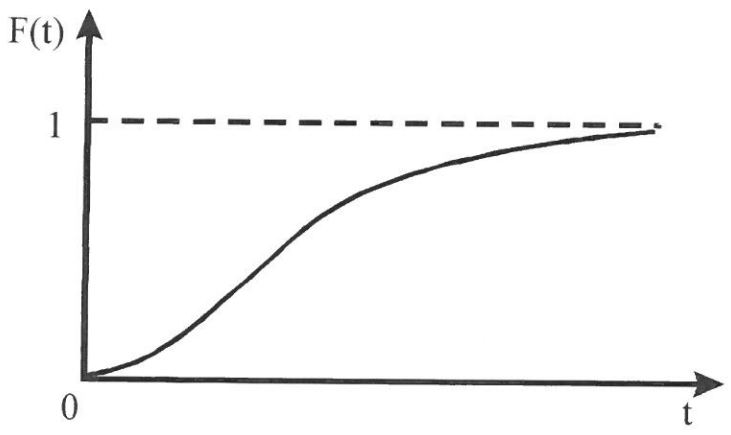
On considère la variable aléatoire T qui mesure la durée de vie ou Temps de Bon Fonctionnement (T.B.F. : Time Between Failure).

1. Fonction de défaillance

La fonction de défaillance est la fonction de répartition F de T.

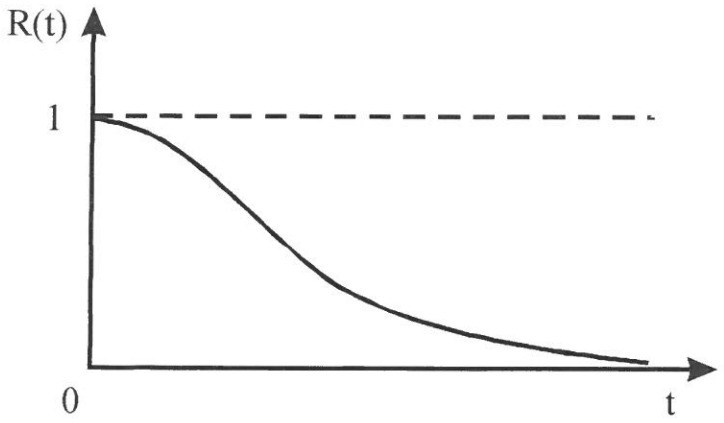
$F(t) = P(T \leq t)$ est la probabilité d'avoir une défaillance avant le temps t

ou encore la probabilité cumulée de défaillance entre 0 et t.



2. Fonction de fiabilité ou de survie

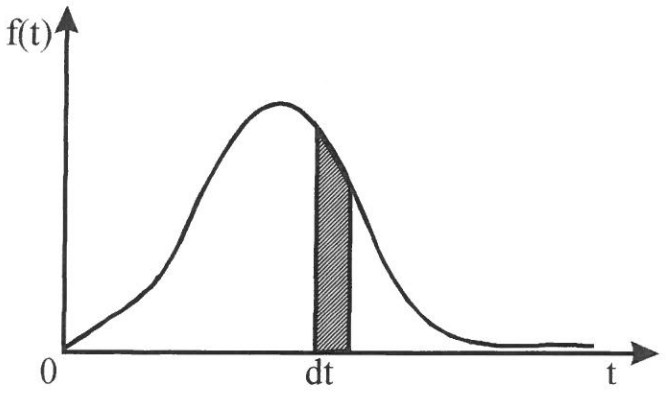
La fonction de fiabilité est la fonction R définie par $R(t) = 1 - F(t)$.
$R(t) = P(T > t)$ est la probabilité d'un fonctionnement sans défaillance pendant la période $[0 ; t]$.



Remarque : Le mot "fiabilité" est la traduction du mot anglais "reliability".

3. Densité de probabilité de F(t)

C'est la fonction f définie par : $f(t) = F'(t) = - R'(t)$.



4. Aspect statistique : estimation de F(t) et de R(t)

Une entreprise veut étudier la fiabilité d'un certain type de matériel. Elle prélève un échantillon de n = 40 éléments de sa production et elle observe la durée de vie, en heures (c'est-à-dire le temps de bon fonctionnement avant la première panne).

Intervalle de temps $]t_{i-1} ; t_i]$ en milliers d'heures	]0 ; 0,5]	]0,5 ; 1]	]1 ; 1,5]	]1,5 ; 2]	]2 ; 3]	]3 ; 4]	]4 ; 6]
Nombres de pannes dans cet intervalle	9	7	5	4	6	5	4

On en déduit :

t_i	500	1 000	1 500	2 000	3 000	4 000	6 000
Nombre n_i d'éléments en panne à l'instant t_i	9	16	21	25	31	36	40
Nombre $N(t_i)$ d'éléments survivants à l'instant t_i	31	24	19	15	9	4	0

Estimations de $F(t_i)$ et de $R(t_i)$ par la méthode des rangs bruts :

On prend $\frac{n_i}{n}$ comme estimation de $F(t_i)$.

t_i	500	1 000	1 500	2 000	3 000	4 000	6 000
$F(t_i)$	0,225	0,4	0,525	0,625	0,775	0,9	1
$R(t_i) = 1 - F(t_i)$							

Estimations de $F(t_i)$ et de $R(t_i)$ par la méthode des rangs moyens

Comme l'échantillon n'est pas très grand, on ne peut conclure à partir de l'observation de cet échantillon que dans la population totale, aucun élément ne survivra plus de 6 000 heures et on ne peut déterminer le nombre de survivants à 6 000 heures.
La méthode des rangs moyens permet de corriger cet écueil.

On prend $\frac{n_i}{n+1}$ comme estimation de $F(t_i)$.

t_i	500	1 000	1 500	2 000	3 000	4 000	6 000
$F(t_i)$	0,219	0,39	0,512	0,61	0,75	0,878	0,976
$R(t_i) = 1 - F(t_i)$							

Estimations de $F(t_i)$ et de $R(t_i)$ par la méthode des rangs médians

Pour corriger d'une autre façon les données quand l'échantillon est petit, on prend $\frac{n_i - 0,3}{n + 0,4}$ comme estimation de $F(t_i)$.

(L'explication de cette formule relève de la théorie des estimateurs, hors programme des S.T.S.)

t_i	500	1 000	1 500	2 000	3 000	4 000	6 000
$F(t_i)$	0,215	0,389	0,52	0,611	0,76	0,884	0,983
$R(t_i) = 1 - F(t_i)$							

Les résultats se situent entre ceux obtenus par les méthodes des rangs bruts et des rangs moyens.

B. Taux d'avarie ou de défaillance :

1. Aspect statistique

Reprenons la situation précédente, on définit le taux d'avarie par unité de temps :

$$\lambda_{]t_{i-1},t_i]} = \frac{\frac{N(t_{i-1}) - N(t_i)}{N(t_{i-1})}}{t_{i-1} - t_i}$$

C'est le quotient du taux de défaillance entre t_{i-1} et t_i par la durée.

On obtient le tableau :

$]t_{i-1}, t_i]$	]0 ; 0,5]	]0,5 ; 1]	]1 ; 1,5]	]1,5 ; 2]	]2 ; 3]	]3 ; 4]	]4 ; 6]
$\lambda_{]t_{i-1},t_i]}$ en taux horaire $\times 10^4$	4,5	4,5	4,2	4,2	4	5,5	5

On voit, en utilisant : $\lambda_{]t_{i-1},t_i]} = \dfrac{\dfrac{N(t_{i-1})}{N(t_0)} - \dfrac{N(t_i)}{N(t_0)}}{t_{i-1} - t_i}$ que : $\lambda_{]t_{i-1},t_i]} = \dfrac{\dfrac{R(t_{i-1}) - R(t_i)}{R(t_{i-1})}}{t_{i-1} - t_i}$.

Ce que l'on peut écrire de façon plus générale : $\lambda_{]t,t+\Delta t]} = \dfrac{R(t) - R(t + \Delta t)}{R(t)\,\Delta t}$

$\lambda_{]t,t+\Delta t]}$ est le taux moyen d'avarie entre t et $t + \Delta t$.

2. Aspect probabiliste

Considérons la limite du taux moyen d'avarie entre t et $t + \Delta t$, défini ci-dessus, lorsque Δt tend vers 0 : $\lambda(t) = \lim\limits_{\Delta t \to 0} \lambda_{]t,t+\Delta t]} = \lim\limits_{\Delta t \to 0} \left[\dfrac{R(t) - R(t + \Delta t)}{\Delta t} \times \dfrac{1}{R(t)}\right]$.

Or on a : $\lim\limits_{\Delta t \to 0} \dfrac{R(t) - R(t + \Delta t)}{\Delta t} = - R'(t) = F'(t) = f(t)$.

Définition

On appelle <u>taux d'avarie instantané à l'instant</u> t, le nombre : $\boxed{\lambda(t) = \dfrac{f(t)}{R(t)} = - \dfrac{R'(t)}{R(t)} = \dfrac{F'(t)}{1 - F(t)}}$

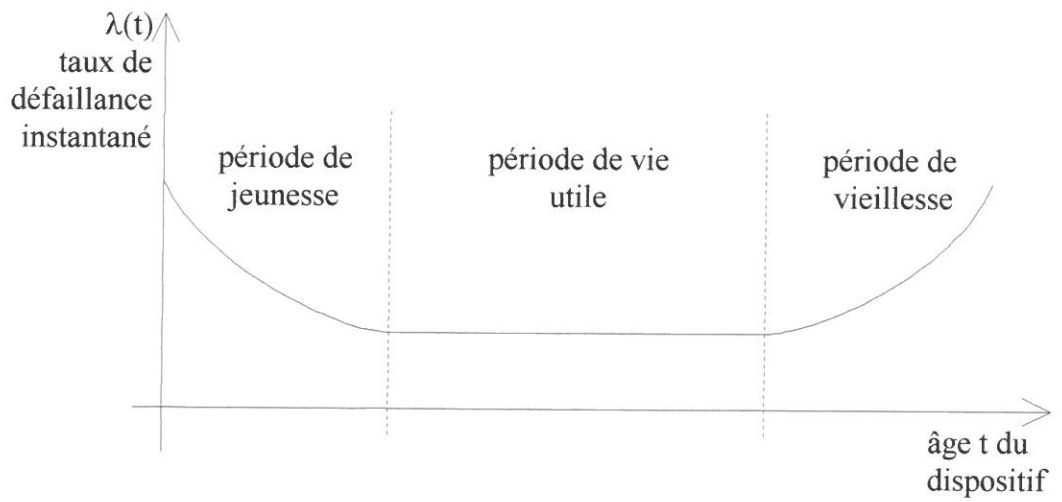
Lorsque Δt est petit, le nombre $\lambda(t)\,\Delta t$ peut être assimilé à la probabilité pour un matériel de tomber en panne pendant l'intervalle de temps $]t ; t + \Delta t]$, sachant qu'il a bien fonctionné jusqu'à l'instant t.

3. Représentation graphique de $\lambda(t)$: courbe en baignoire

La courbe, dite "courbe en baignoire", donne l'évolution du taux de défaillance instantané $\lambda(t)$ en fonction de l'âge t du matériel. On distingue trois périodes qui décrivent trois catégories de défaillances différentes suivant l'âge du dispositif :

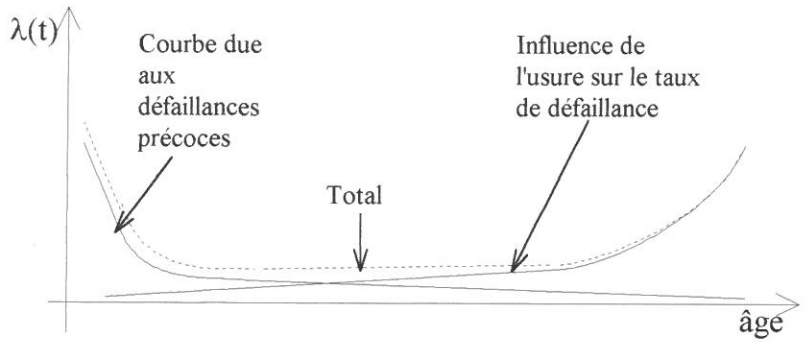
- la <u>période de jeunesse</u> (ou période de mortalité infantile ou période des défaillances précoces) pendant laquelle le taux de défaillance décroît ; en électronique, on essaie de s'affranchir de cette période par le déverminage et en mécanique par la période de rodage ou d'essais non destructifs ;
- la <u>période de vie utile</u> qui correspond à la maturité du dispositif durant laquelle les défaillances sont "totalement aléatoires" et le taux de défaillance constant ;
- la <u>période de vieillesse</u> pendant laquelle le taux de défaillance croît.

Ces trois périodes sont schématisées par la célèbre "courbe en baignoire" :

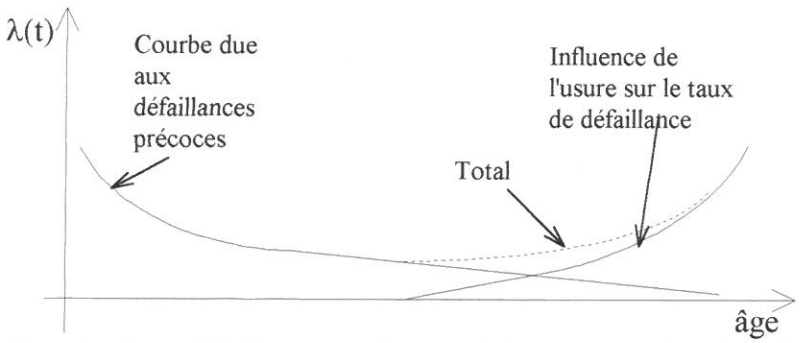


Ces trois phases de vie d'un dispositif sont communes aux dispositifs mécaniques et électroniques, mais n'ont pas la même durée. De plus, la période à taux de défaillance constant ne serait que la juxtaposition d'une période de défaillance de jeunesse qui s'éternise avec une période de dégradation qui s'établit progressivement. La dégradation qui s'établit apparaît très lentement et très progressivement dès le début pour l'électronique, beaucoup plus tard, mais plus fortement pour la mécanique. La période d'usure en électronique est extrêmement éloignée. On suppose qu'elle existe effectivement pour tous les types de composants, mais elle n'a été constatée en fait que sur un nombre restreint de types de composants, probablement parce que les composants électroniques sont rapidement obsolescents.

L'emploi du terme "usure" est mal approprié mais usuel. Il serait plus exact de parler de dégradation car l'usure est un phénomène physique caractérisé par le frottement de deux corps en contact.



Combinaison défaillances précoces et d'usure pour l'électronique



Combinaison défaillances précoces et d'usure pour la mécanique

4. Caractérisation de la fiabilité à partir du taux de défaillance

$\lambda(t) = -\dfrac{R'(t)}{R(t)}$ pour $t > 0$.

D'où $\int_0^t \lambda(u)\,du = \left[-\ln R(u)\right]_0^t = -\ln R(t)$ (car $R(0) = 1$)

D'où $\boxed{R(t) = e^{-\int_0^t \lambda(u)\,du}}$

C. La M.T.B.F. (Mean Time Between Failure)

(traduit par "Moyenne des Temps de Bon Fonctionnement".)

C'est l'espérance mathématique de la variable aléatoire T continue, définie sur [0 ; + ∞[et de densité de probabilité f : $\boxed{M.T.B.F. = E(T) = \int_0^{+\infty} t\,f(t)\,dt.}$

Avec une intégration par parties : $\boxed{M.T.B.F. = \int_0^{+\infty} R(t)\,dt.}$

D. Montage en série, en parallèle

1. Montage en série

Lorsqu'un système est constitué de composants montés en série, le système est défaillant dès qu'un des composants est en panne. Le système ne fonctionne donc que lorsque tous les composants fonctionnent.

On peut généralement considérer que les défaillances des différents composants sont indépendantes ; la probabilité que le système fonctionne à un instant t est donc le produit des probabilités de bon fonctionnement de chacun des composants.

Si l'on a n composants montés en série, de fonctions de fiabilité respectives R_1, R_2,... R_n, la fonction de fiabilité du système est définie par :

$$R(t) = P(T > t) = P[(T_1 > t) \cap (T_2 > t) \cap ... \cap (T_n > t)]$$
$$= P(T_1 > t) \times P(T_2 > t) \times ... \times P(T_n > t)$$

$$\boxed{R(t) = R_1(t) \times R_2(t) \times ... \times R_n(t).}$$

2. Montage en parallèle

Dans le cas où tous les composants sont montés en parallèle, le système est défaillant lorsque tous les composants sont défaillants.

$$F(t) = P(T \leq t) = P[(T_1 \leq t) \cap (T_2 \leq t) \cap ... \cap (T_n \leq t)]$$
$$= F_1(t) \times F_2(t) \times ... \times F_n(t)$$

Soit $1 - R(t) = (1 - R_1(t)) \times (1 - R_2(t)) \times ... \times (1 - R_n(t))$

$$\boxed{R(t) = 1 - (1 - R_1(t)) \times (1 - R_2(t)) \times ... \times (1 - R_n(t))}$$

E. Un exemple purement mathématique

1. Soit la fonction $f : [0 ; +\infty[\longrightarrow \mathbb{R}$

$$t \longmapsto \frac{0{,}02\ t}{(0{,}01\ t^2 + 1)}$$

Montrer que f peut être considérée comme la densité de probabilité d'une variable aléatoire T, c'est-à-dire que : $\begin{cases} \forall\ t \in [0,+\infty] \quad f(t) \geq 0 \\ \int_0^{+\infty} f(t)\ dt = 1 \end{cases}$

2. En supposant que T représente le temps de bon fonctionnement d'un matériel, déterminer la fonction de défaillance F, la fonction de fiabilité R et le taux d'avarie λ.

Représenter graphiquement R et λ.

3. Calculer la M.T.B.F.

Corrigé

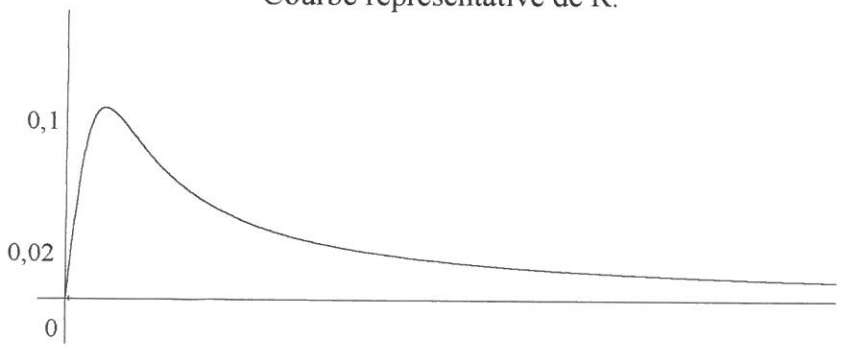
1. $\forall\ t \in [0,+\infty] \quad f(t) \geq 0$

$$\int_0^{+\infty} f(t)\ dt = \lim_{A \to 0} \left[\frac{-1}{0{,}01\ t^2 + 1} \right]_0^A = 1.$$

2. $F(t) = \int_0^t f(x)\ dx = \left[\frac{-1}{0{,}01\ x^2 + 1} \right]_0^t = 1 - \frac{1}{0{,}01\ t^2 + 1}.$

$$R(t) = 1 - F(t) = \frac{1}{0{,}01\ t^2 + 1} \quad \text{et} \quad \lambda(t) = \frac{f(t)}{R(t)} = \frac{0{,}02\ t}{0{,}01\ t^2 + 1}.$$

Courbe représentative de R.



Courbe représentative de λ.

3. M.T.B.F. $= \int_0^{+\infty} \frac{dt}{0{,}01\,t^2+1} = \int_0^{+\infty} \frac{10\,du}{u^2+1} = 10\,\left[\text{Arctan } u\right]_0^{+\infty} = 5\pi \approx 15{,}7.$

V. Loi exponentielle

A. Généralités

Pour certains composants, comme les composants électroniques, le taux d'avarie est sensiblement constant. C'est aussi le cas pour la plupart des matériels pendant leur vie utile.

Un taux d'avarie constant caractérise une loi de fiabilité exponentielle. On a en effet immédiatement, si $\forall\, t,\ \lambda(t) = \lambda$:

Fonction de fiabilité : $R(t) = e^{-\lambda t}$;

Fonction de défaillance : $F(t) = 1 - e^{-\lambda t}$;

Densité de défaillance : $f(t) = \lambda\, e^{-\lambda t}$.

B. M.T.B.F. et écart type

M.T.B.F. $= \int_0^{+\infty} R(t)\,dt = \int_0^{+\infty} e^{-\lambda t}\,dt = \left[\frac{e^{-\lambda t}}{-\lambda}\right]_0^t$ D'où $\boxed{\text{M.T.B.F.} = \frac{1}{\lambda}}$

La variance de T est $\text{Var}(T) = E(T^2) - [E(T)]^2 = \int_0^{+\infty} t^2\, f(t)\,dt$ $- (\text{M.T.B.F.})^2$

Par une double intégration par parties, on trouve :

$\int_0^{+\infty} t^2\, f(t)\,dt = \lambda \int_0^{+\infty} t^2\, e^{-\lambda t}\,dt = \frac{2}{\lambda^2}$; d'où $\text{Var}(T) = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$.

Ainsi, l'écart type de T est $\boxed{\sigma(T) = \frac{1}{\lambda} = \text{M.T.B.F.}}$

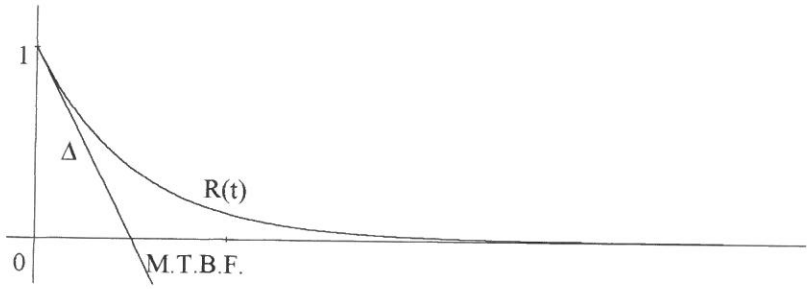
Une remarque intéressante

$R(\text{M.T.B.F.}) = R\left(\frac{1}{\lambda}\right) = e^{-\lambda(1/\lambda)} = e^{-1} \approx 0{,}368.$

La M.T.B.F. est donc l'abscisse du point de la courbe représentative de la fonction R qui a comme ordonnée $e^{-1} \approx 0{,}368$.

C. Recherche graphique de la M.T.B.F.

La tangente Δ à la courbe représentative de la fonction R au point d'abscisse 0 a pour coefficient directeur : $R'(0) = -f(0) = -\lambda$. Elle a pour équation $y = -\lambda t + 1$.



Les cordonnées (t ; y) du point d'intersection de Δ avec l'axe des abcsisses vérifient donc le système : $\begin{cases} y = -\lambda t + 1 \\ y = 0 \end{cases}$

Ainsi, $\boxed{t = \frac{1}{\lambda} = \text{M.T.B.F.}}$

D. Utilisation du papier semi-logarithmique

Le papier semi-logarithmique est gradué selon une échelle arithmétique sur l'axe des abscisses et selon une échelle logarithmique sur l'axe des ordonnées.

Intérêt

La représentation graphique d'une fonction exponentielle est une droite.

Sur un exemple

a). On prend $\lambda = 0{,}02$.

Calculer R(t) pour les valeurs de t : 0 ; 25 ; 50 ; 75 ; 100 ; 125 et 150.

Placer alors les points obtenus sur la feuille de papier semi-logarithmique.

Vérifier l'alignement de ces points.

b). On a $\text{M.T.B.F.} = \frac{1}{\lambda} = 50$.

Vérifier que cette M.T.B.F. correspond à l'abscisse du point d'ordonnée $e^{-1} \approx 0{,}368$.

On retrouve ainsi graphiquement la M.T.B.F.

A PROPOS DE SUJETS D'EXAMEN

Présentation de quelques sujets d'examen

La préoccupation du professeur de mathématiques enseignant en section de technicien supérieur, qui conditionne la formation qu'il donne à ses élèves, est de prévoir quel type de sujet sera proposé à l'examen.

Pour les professeurs qui participent aux commissions chargées de l'élaboration de ces sujets de Brevets de Techniciens Supérieurs, la question est de savoir "jusqu'où peut-on aller ?".

Pour tenter d'aider à trouver une réponse à ces questions, une sélection de six exercices, inspirés de sujets posés à divers B.T.S., est proposée. Ce choix repose sur deux critères essentiels : d'une part, la conformité avec les programmes, dans la lettre et dans l'esprit, en particulier la rédaction du texte, et d'autre part, les difficultés qui paraissent raisonnables pour des exercices à traiter en une heure, présentation de la copie comprise.

Liste des travaux pratiques des modules de programme abordés dans ces exercices :

Statistique descriptive 1

T.P. 1 : Etude de séries statistiques à une variable.

Calcul des probabilités 1 & 2

T.P. 1 : Emploi de dénombrements pour le calcul de probabilités.

T.P. 2 : Exemples d'étude de situations de probabilités faisant intervenir des variables aléatoires suivant une loi binomiale, de Poisson ou normale.

Statistique inférentielle 1

T.P. 1 : Exemples d'estimation d'une moyenne ou d'un pourcentage par un intervalle de confiance.

T.P. 2 : Exemples de construction et d'utilisation d'un test de validité d'hypothèse relatif à une moyenne ou à un pourcentage.

Comptabilité et Gestion 1990 :

La troisième partie de l'exercice concerne un test de validité d'hypothèse relatif à une moyenne. La rédaction des questions prend en compte le libellé du programme «*Exemple de ...*», ce qui précise bien que les élèves doivent être guidés dans ce type d'exercice.

Comptabilité et Gestion 1993 & Mécanique et Automatismes Industriels 1990 :

Même remarque à propos de la deuxième question.

Informatique de Gestion 1989 (Nouméa) :

À noter l'introduction de la partie B : il est fait allusion à l'ajustement analytique d'une série empirique à une loi normale, mais les élèves, qui n'en auraient jamais entendu parlé (malheureusement), ne sont pas pénalisés pour autant. À remarquer également, la distinction très nette entre partie statistique et probabilités.

Informatique de Gestion 1992 (Nouméa) & Comptabilité et Gestion 1993 (Nouméa) :

L'introduction du premier de ces deux sujets explique parfaitement l'aspect "fréquentiste" de cette situation : après avoir effectué une observation statistique, en l'absence d'autre information, on assimile les fréquences observées à des probabilités, la variable statistique devient aléatoire.

Dans ces deux exercices, les dernières questions nécessitent l'utilisation de la correction de continuité : la rédaction est telle, qu'un candidat n'ayant pas été formé à cette technique (malheureusement), peut tout au plus être surpris par la formulation de la question, mais peut y répondre sans difficulté.

Sujets de B.T.S.

COMPTABILITE ET GESTION ✧ Session 1990

Le but de cet exercice est d'étudier des situations de probabilités faisant intervenir des variables aléatoires suivant une loi normale (première partie) ou une loi binomiale (deuxième partie), puis de construire et d'utiliser un test de validité d'hypothèse relatif à une moyenne (troisième partie).

Dans cet exercice, pour les résultats numériques demandés, on donnera les valeurs décimales arrondies à 10^{-3} près.

Première partie:

On considère que la variable aléatoire X mesurant le chiffre d'affaires journalier d'un hypermarché suit la loi normale de moyenne 1,5 million de francs et d'écart type 0,3 million de francs.

1°. Calculer la probabilité que X prenne une valeur inférieure à 1,8 million de francs.

2°. Calculer la probabilité que X prenne une valeur supérieure à 2 millions de francs.

Deuxième partie:

On s'intéresse aux chiffres d'affaires journaliers de cet hypermarché pendant trente jours ouvrables et on suppose que les trente tirages d'un chiffre d'affaires journalier sont indépendants. Soit Y la variable aléatoire mesurant ainsi sur trente jours ouvrables, le nombre de jours où le chiffre d'affaires est supérieur à 2 millions de francs.

1°. Expliquer pourquoi Y suit une loi binomiale et donner les paramètres de cette loi binomiale.

2°. Calculer la probabilité de l'événement $Y \geq 1$. (Pour ce calcul, on utilisera la valeur décimale arrondie à 10^{-3} près obtenue à partir de la première partie.)

Troisième partie:

Avant d'engager une campagne publicitaire, la direction de l'hypermarché vous demande de construire un test unilatéral qui, au vu des chiffres d'affaires journaliers des trente jours ouvrables suivant cette campagne, permettra de décider si, au seuil de signification 5 %, la moyenne des chiffres d'affaires journaliers a augmenté, c'est-à-dire dépassé 1,5 million de francs, à la suite de cette campagne publicitaire.

1°. Construction du test unilatéral :

On note m la moyenne *inconnue* de la nouvelle population des chiffres d'affaires journaliers obtenus après la campagne publicitaire et on suppose que l'écart type de cette population est encore de 0,3 million de francs. Soit Z la variable aléatoire qui, à tout échantillon aléatoire, non exhaustif, de trente chiffres d'affaires journaliers de cette nouvelle population, associe la moyenne de ceux-ci.

On suppose que Z suit la loi normale de moyenne m et d'écart type $\frac{0,3}{\sqrt{30}}$.

a). Choisir une hypothèse nulle H_0 et une hypothèse alternative H_1 pour ce test *unilatéral* .

b). Déterminer le nombre réel h tel que, sous l'hypothèse H_0 , on ait $P(Z \leq h) = 0,95$.

c). Enoncer la règle de décision de ce test.

2°. Utilisation de ce test :

Les chiffres d'affaires journaliers pendant les trente jours ouvrables suivant la campagne publicitaire sont donnés par le tableau suivant :

Chiffre d'affaires journalier (en millions de F)	1,2	1,3	1,4	1,5	1,6	1,7	1,8	1,9	2	2,1	2,2	2,3
Nombre de jours	1	2	2	8	8	2	2	1	2	1	0	1

a). Calculer la moyenne des chiffres d'affaires journaliers pendant ces trente jours.

b). En appliquant la règle de décision du test à cet échantillon de trente chiffres d'affaires journaliers que l'on assimile à un échantillon aléatoire non exhaustif, peut-on conclure, au seuil de signification 5 % qu'à la suite de la campagne publicitaire la moyenne des chiffres d'affaires journaliers a dépassé 1,5 millions de francs ?

CORRIGE

Première partie:

1°. La variable aléatoire X mesurant le chiffre d'affaires journalier de l'hypermarché considéré suit la loi normale N(1,5 ; 0,3). La probabilité cherchée est calculée à l'aide de la variable aléatoire centrée réduite $T = \frac{X - 1,5}{0,3}$ associée à X, et qui suit donc la loi normale N(0,1).

$$P(X < 1,8) = P\left(\frac{X - 1,5}{0,3} < \frac{1,8 - 1,5}{0,3}\right) = P(T < 1) = P(T \leq 1) = \pi(1) \approx 0,841 \quad (\text{à } 10^{-3} \text{ près}).$$

2°. De même : $P(X>2) = P\left(T > \frac{0,5}{0,3}\right) = 1 - P\left(T \leq \frac{0,5}{0,3}\right) = 1 - \pi(1,666...) \approx 0,048$ (à 10^{-3} près).

Deuxième partie :

1°. En tirant au hasard un chiffre d'affaires journalier, l'un des deux événements suivants est réalisé :

* le chiffre d'affaires est supérieur à deux millions de francs,
  (événement de probabilité 0,048 d'après la première partie)
* le chiffre d'affaires n'est pas supérieur à deux millions de francs,
  (événement de probabilité 1 - 0,048 = 0,952).

Puisque l'on suppose que ces trente tirages constituent des événements indépendants, la variable aléatoire Y mesurant le nombre de ces tirages pour lesquels le chiffre d'affaires est supérieur à deux millions de francs suit la loi binomiale de paramètres $n = 30$ et $p = 0{,}048$.

2°. $P(Y \geq 1) = 1 - P(Y < 1) = 1 - P(Y = 0) = 1 - C_{30}^{0}\, 0{,}952^{30} \approx 0{,}771$ (à 10^{-3} près).

Troisième partie :

1°. a). Choix de H_0 : la moyenne des chiffres d'affaires journaliers de l'hypermarché après la campagne publicitaire est $\mu = 1{,}5$ (million de francs).

Choix de H_1 : $\mu > 1{,}5$.

b). Sous l'hypothèse H_0, on a $\mu = 1{,}5$, donc Z suit la loi normale $N(1{,}5 ; 0{,}3)$ et la variable aléatoire centrée réduite V associée à Z est définie par $V = \frac{\sqrt{30}}{0{,}3}(Z - 1{,}5)$.

Par suite, $P(Z \leq h) = 0{,}95$ équivaut successivement à :

$$P\left(V \leq \frac{\sqrt{30}}{0{,}3}(h - 1{,}5)\right) = 0{,}95 \qquad \pi\left(\frac{\sqrt{30}}{0{,}3}(h - 1{,}5)\right) = 0{,}95 \qquad \frac{\sqrt{30}}{0{,}3}(h - 1{,}5) = 1{,}645$$

$$h = \frac{\sqrt{30}}{0{,}3} \times 1{,}645 + 1{,}5 \qquad \text{d'où} \qquad h = 1{,}590.$$

c). Enoncé de la règle de décision :

On prélève un échantillon non exhaustif de taille 30 dans la population des chiffres d'affaires journaliers obtenus après la campagne publicitaire. On calcule la moyenne m de cet échantillon.

Si $m \leq 1{,}590$: on accepte H_0 .

Si $m > 1{,}590$: on rejette H_0 .

2°. a). $m \approx 1{,}623$.

b). $m > 1{,}590$, on rejette H_0 .

On conclut, au seuil de signification 5 %, qu'à la suite de la campagne publicitaire la moyenne des chiffres d'affaires journaliers a augmenté, c'est-à-dire dépassé 1,5 million de francs.

COMPTABILITE ET GESTION ✧ Session 1993

Dans cet exercice : pour les valeurs numériques, on donnera les approximations décimales arrondies à 10^{-2} près. Les questions **1°.** et **2°.** sont indépendantes.

Dans un centre de renseignements téléphoniques, une étude statistique a été réalisée sur le temps d'attente, exprimé en secondes, subi par la clientèle avant d'amorcer la conversation avec un employé. Les résultats de cette étude conduisent à supposer que la variable aléatoire X qui associe à tout client le temps d'attente qu'il subit, suit la loi normale de moyenne 18 et d'écart type 7,2.

1°. a). Calculer la probabilité que lors d'un appel au centre, un client :

* n'ait à subir aucune attente (c'est-à-dire $P(X \leq 0)$).
* ait à subir une attente de plus de 20 secondes.

b). On imagine qu'au cours d'une certaine semaine, un même client doit donner au centre, cinq appels, indépendants les uns des autres. On note Y la variable aléatoire exprimant le nombre de fois où, au cours de ces cinq appels, le temps d'attente est supérieur à 20 secondes. Préciser la loi de probabilité suivie par Y et donner ses paramètres.

Calculer $P(Y = 2)$ et $P(Y \geq 1)$.

2°. Dans le but de diminuer le temps d'attente, une restructuration des services du centre de renseignements téléphoniques est réalisée. Pour tester cette restructuration, on effectue une enquête sur un échantillon de 100 clients.

a). Soit $\overline{X}$ la variable aléatoire mesurant le temps d'attente moyen exprimé en secondes par échantillon non exhaustif de 100 clients.

La variable aléatoire $\overline{X}$ suit approximativement la loi $N\left(18, \frac{7,2}{\sqrt{100}}\right)$.

On se propose de construire un test unilatéral permettant de décider, au risque 1 %, si le temps d'attente moyen a été réduit par la restructuration. Pour répondre à cette question :

- Choisir une hypothèse nulle H_0 et une hypothèse alternative H_1.
- Déterminer la région critique au seuil de signification 1 %.
- Enoncer la règle de décision.

b). Voici les résultats d'une enquête réalisée auprès de 100 clients.

Temps d'attente (exprimé en secondes)	[0,5[	[5,10[	[10,20[	[20,25[	[25,30[	[30,35[	[35,40[
Nombre de clients	10	16	24	24	12	10	4

Calculer la moyenne des temps d'attente pour cet échantillon.

c). Utiliser le test avec cet échantillon pour argumenter votre conclusion.

MECANIQUE ET AUTOMATISMES INDUSTRIELS ✧ Session 1990

Une usine fabrique des engrenages, et on désigne par X la variable aléatoire mesurant le diamètre, exprimé en millimètres, de ces engrenages. On admet que X suit une loi normale de paramètres M et σ.

1°. Dans cette question, on suppose que l'on a : M = 23,65 (mm) et σ = 0,02 (mm).

a). Un engrenage est utilisable lorsque la mesure de son diamètre, exprimée en millimètres, appartient à l'intervalle [23,61 ; 23,70].

Calculer la probabilité qu'un engrenage pris au hasard dans la production de l'usine soit utilisable.

b). Déterminer un intervalle de centre M = 23,65 tel qu'un engrenage tiré au hasard dans la production de l'usine ait un diamètre dont la mesure appartienne à cet intervalle avec la probabilité 0,9.

2°. Un client commande un lot d'engrenages, dont on lui annonce que la moyenne des mesures des diamètres est 23,65 mm. Ce client veut vérifier cette affirmation et mesure les diamètres de 100 engrenages . Il obtient les résultats suivants:

Diamètre	[23,59 ; 23,61[	[23,61 ; 23,63[	[23,63 ; 23,65[	[23,65 ; 23,67[	[23,67 ; 23,69[
Effectifs	6	8	51	30	5

a) Calculer la moyenne et l'écart type de cet échantillon.

b) Le client considère que l'écart type de cet échantillon est une bonne approximation de l'écart type du lot qu'il a reçu. Peut-il admettre au risque de 5 % l'affirmation de son fournisseur ?

Pour répondre à cette question **:**

I . Construire un test de validité d'hypothèse.

- Choisir une hypothèse nulle H_0 et une hypothèse alternative H_1.
- Déterminer la région critique au seuil de signification 5 %.
- Enoncer la règle de décision.

II . Utiliser le test avec l'échantillon.

CORRIGE

1°. a). Soit T la variable aléatoire centrée réduite associée à X.

On a $T=\dfrac{X - 23,65}{0,02}$, T suit la loi N (0,1), d'où le calcul de $P(23,61 \le X \le 23,70)$:

$$P\left(\frac{23,61 - 23,65}{0,02} \le T \le \frac{23,70 - 23,65}{0,02},\right) = P(-2 \le T \le 2,5) = P(T \le 2,5) - P(T \le -2) = \pi(2,5) - \pi(-2).$$

Pour tout nombre réel t, $\pi(-t) = 1 - \pi(t)$, donc :

$$P(23,61 \le X \le 23,70) = \pi(2,5) + \pi(2) - 1 = 0,9938 + 0,9772 - 1 \approx 0,971.$$

b) Il s'agit de déterminer un intervalle [M - h , M + h] où M = 23,65 et h est un nombre réel tel que :

$$P(M - h \leq X \leq M + h) = 0,9.$$

Cette égalité est successivement équivalente à : $P(23,65-h \leq X \leq 23,65+h) = 0,9$

$$P\left(\frac{-h}{0,02} \leq T \leq f(h;0,02)\right) = 0,9 \quad ; \quad 2\pi\left(\frac{h}{0,02}\right) - 1 = 0,9 \quad ; \quad \pi\left(\frac{h}{0,02}\right) = 0,95.$$

Par lecture inverse de la table du formulaire on a : $\frac{h}{0,02} = 1,645$ d'où $h = 0,0329$.

L'intervalle cherché est [23,617 ; 23,683].

2° a) Pour le calcul de la moyenne et de l'écart type de l'échantillon, on utilise les centres des classes:

Centre de la classe	23,60	23,62	23,64	23,66	23,68
Effectifs	6	8	51	30	5

La calculatrice donne la moyenne de cet échantillon : 23,644 (mm) et l'écart type de cet échantillon: 0,0177 (mm).

b). I. ***Construction du test de validité d'hypothèse relatif à la moyenne.***

- **Choix de H_0 :** l'affirmation du fournisseur est correcte, la moyenne des mesures des diamètres des engrenages est $\mu = 23,65$.

Le choix de H_1 s'impose alors, le test étant bilatéral : $\mu \neq 23,65$.

- **Détermination d'une région critique au seuil 5 % :**

Soit $\overline{X}$ la variable aléatoire qui associe, à tout échantillon non exhaustif de taille $n = 100$, la moyenne des mesures des diamètres des engrenages de cet échantillon. Puisque le client considère que l'écart type de cet échantillon est une bonne approximation de l'écart type du lot qu'il a reçu, on suppose que l'écart type de la population est $\sigma = 0,0177$.

On sait que $\overline{X}$ suit la loi normale $N\left(m, \frac{\sigma}{\sqrt{n}}\right)$, soit la loi $N(23,65 ; 0,0018)$.

Pour tout $t > 0$, $\quad P\left(\mu - \frac{\sigma}{\sqrt{n}} \leq \overline{X} \leq \mu + \frac{\sigma}{\sqrt{n}}\right) = 2\pi(t) - 1.$

Pour $2\pi(t) - 1 = 0,95$, la table donne $t = 1,96$.

On en déduit : $P(23,65 - 1,96 \times 0,0018 \leq \overline{X} \leq 23,65 + 1,96 \times 0,0018) = 0,95.$

Le calcul de valeurs approchées à 10^{-3} près par défaut pour la borne inférieure et par excès pour la borne supérieure donne alors : $P(23,646 \leq \overline{X} \leq 23,654) = 0,95.$

Si H_0 est vraie on a 95 % de chances de prélever un échantillon aléatoire non exhaustif de taille $n = 100$, dont la moyenne appartient à l'intervalle $I = [23,646 ; 23,654]$.

• **Règle de décision:**

Soit $\overline{X}$ la moyenne des mesures des diamètres des engrenages d'un échantillon aléatoire non exhaustif de taille n = 100.

Si $\overline{X} \in I$, on accepte H_0 et si $\overline{X} \notin I$ on rejette H_0.

II. ***Utilisation du test avec l'échantillon étudié précédemment:***

La moyenne de cet échantillon est 23,644 mm. Puisque $23{,}644 \notin I$, on rejette H_0. On considère que l'affirmation du fournisseur est incorrecte, au seuil de signification 5 % les pièces du lot dont est tiré l'échantillon n'ont pas un diamètre moyen de 23,65 mm.

INFORMATIQUE DE GESTION ✧ Session 1989 Nouméa

Une station service d'une grande surface a relevé pendant une semaine, la demande exprimée en litres de chacun de ses clients :

Demande en litres	[5,15[	[15,20[	[20,25[	[25,30[	[30,35[	[35,40[	[40,50[	[50,60[
Nombre de clients	11	45	158	223	273	132	44	4

A. Etude statistique.

1. Représenter l'histogramme de cette série dans un repère orthogonal $(O ; \vec{i}, \vec{j})$. On prend pour unités graphiques : • 1 cm pour 5 litres sur l'axe des abscisses

• 5 cm pour 100 clients sur l'axe des ordonnées.

2. Calculer à 10^{-3} près la moyenne $\overline{x}$ et l'écart type σ_x de cette série statistique. Les résultats intermédiaires ne sont pas demandés.

3. Calculer à 10^{-1} près les pourcentages respectifs de clients dont la demande, exprimée en litres, est située entre $\overline{x} - \sigma_x$ et $\overline{x} + \sigma_x$, entre $\overline{x} - 2\sigma_x$ et $\overline{x} + 2\sigma_x$, entre $\overline{x} - 3\sigma_x$ et $\overline{x} + 3\sigma_x$.

B. Etude probabiliste.

Soit X la variable aléatoire mesurant la demande, exprimée en litres, d'un client au cours d'une semaine donnée. Les résultats précédents permettent de convenir que X suit la loi normale de moyenne m = 30 et d'écart type $\sigma = 7$.

Les résultats de cette partie seront donnés à 10^{-2} près.

1. Calculer la probabilité que X prenne une valeur comprise strictement entre 25 et 35 .

2. Calculer la probabilité que X prenne une valeur supérieure ou égale à 40.

C. Campagne publicitaire et probabilité.

Dans un but publicitaire, la direction de la grande surface a décidé de donner un cadeau à tout client dont la demande est au moins de 40 litres au cours de la semaine.

On suppose que les demandes des clients constituent des événements mutuellement indépendants.

1. Soient Y la variable aléatoire mesurant le nombre de clients ayant droit à un cadeau parmi les cent vingt premiers consommateurs se présentant à la station service, et un entier naturel k, $0 \leq k \leq 120$. Calculer la probabilité de l'événement $Y = k$.

2. On décide d'approcher la loi de probabilité de la variable aléatoire Y par une loi de Poisson de paramètre $\lambda = 9$. Calculer à 10^{-3} près, en utilisant cette approximation, la probabilité de chacun des événements : **a).** $Y = 10$ et **b).** $Y \geq 15$.

INFORMATIQUE DE GESTION ✧ Session 1992 Nouméa

Dans la revue *Science et Vie*, Economie *Magazine*, - hors-série 90 / 91, on peut lire : "On estime à 60,5 % le pourcentage de Français partant au moins une fois en vacances dans le courant de l'année". On admet alors que la probabilité qu'une personne prise au hasard dans la population parte en vacances dans le courant de l'année est 0,605. On considère 100 personnes prises au hasard avec remise parmi la population française.

1°. On désigne par X la variable aléatoire mesurant, parmi ces 100 personnes, le nombre de celles qui ne partent pas en vacances dans le courant de l'année.

a). Justifier que la loi de probabilité suivie par la variable aléatoire X est une loi binomiale dont on précisera les paramètres.

b). Calculer l'espérance mathématique et l'écart type de la variable aléatoire X.

c). Calculer la probabilité de l'événement "X = 45". Pour ce calcul on prendra $C_{100}^{45} = 6{,}145 \times 10^{28}$.

2°. On décide d'approcher la loi de la variable aléatoire discrète X par la loi normale de paramètres $m = 39{,}5$ et $\sigma = 4{,}89$. On note Y une variable aléatoire suivant la loi N(39,5 ; 4,89). En utilisant cette approximation, calculer :

a). la probabilité que 45 personnes exactement parmi les 100 ne partent pas en vacances dans le courant de l'année, c'est-à-dire : $P(44{,}5 \leq Y \leq 45{,}5)$;

b). la probabilité qu'au plus 30 de ces 100 personnes ne partent pas en vacances dans le courant de l'année, c'est-à-dire : $P(Y \leq 30{,}5)$.

COMPTABILITE ET GESTION ✧ Session 1993 Nouméa

Une entreprise assure la production de trois objets A_1, A_2 et A_3, en quantités hebdomadaires respectives x_1, x_2 et x_3. Un "programme de production" (hebdomadaire) s'exprime par un vecteur $\vec{x} = (x_1, x_2, x_3)$ de $\mathbb{R}^3$.

(Par exemple, le programme de production (20, 10, 30) correspond, pour la semaine concernée, à une production de 20 objets A_1 , 10 objets A_2 et 30 objets A_3)

1. Calcul vectoriel :

Pour réaliser un programme de production $\vec{x} = (x_1, x_2, x_3)$, on utilise y_1 kilogrammes de matière première, y_2 heures de travail et y_3 kilowattheures d'énergie, ce que l'on représente par le vecteur $\vec{y} = (y_1, y_2, y_3)$ de $\mathbb{R}^3$. On sait que l'application h, de $\mathbb{R}^3$ dans $\mathbb{R}^3$, qui à $\vec{x}$ associe $\vec{y}$ est linéaire.

Sa matrice associée, dans la base canonique de $\mathbb{R}^3$, est $H = \begin{pmatrix} 2 & 4 & 1 \\ 1 & 2 & 2 \\ 1 & 1 & 1 \end{pmatrix}$. Ainsi $\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 2 & 4 & 1 \\ 1 & 2 & 2 \\ 1 & 1 & 1 \end{pmatrix} \times \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$.

a). Calculer le vecteur $\vec{Y}$ associé au vecteur $\vec{X} = (20, 10, 30)$.

b). Déterminer le vecteur $\vec{X}$ qui a pour vecteur associé $\vec{Y} = (230, 130, 100)$.

2. Calcul de probabilités :

Une étude a permis de constater, compte tenu des contraintes extérieures, que, quel que soit le programme de production hebdomadaire choisi, la probabilité de le réaliser effectivement est égale à 0,7. (On suppose que les réalisations des programmes hebdomadaires de production sont indépendantes les unes des autres). On note V la variable aléatoire qui mesure le nombre de programmes de production hebdomadaires réalisés, par période de 50 semaines de production.

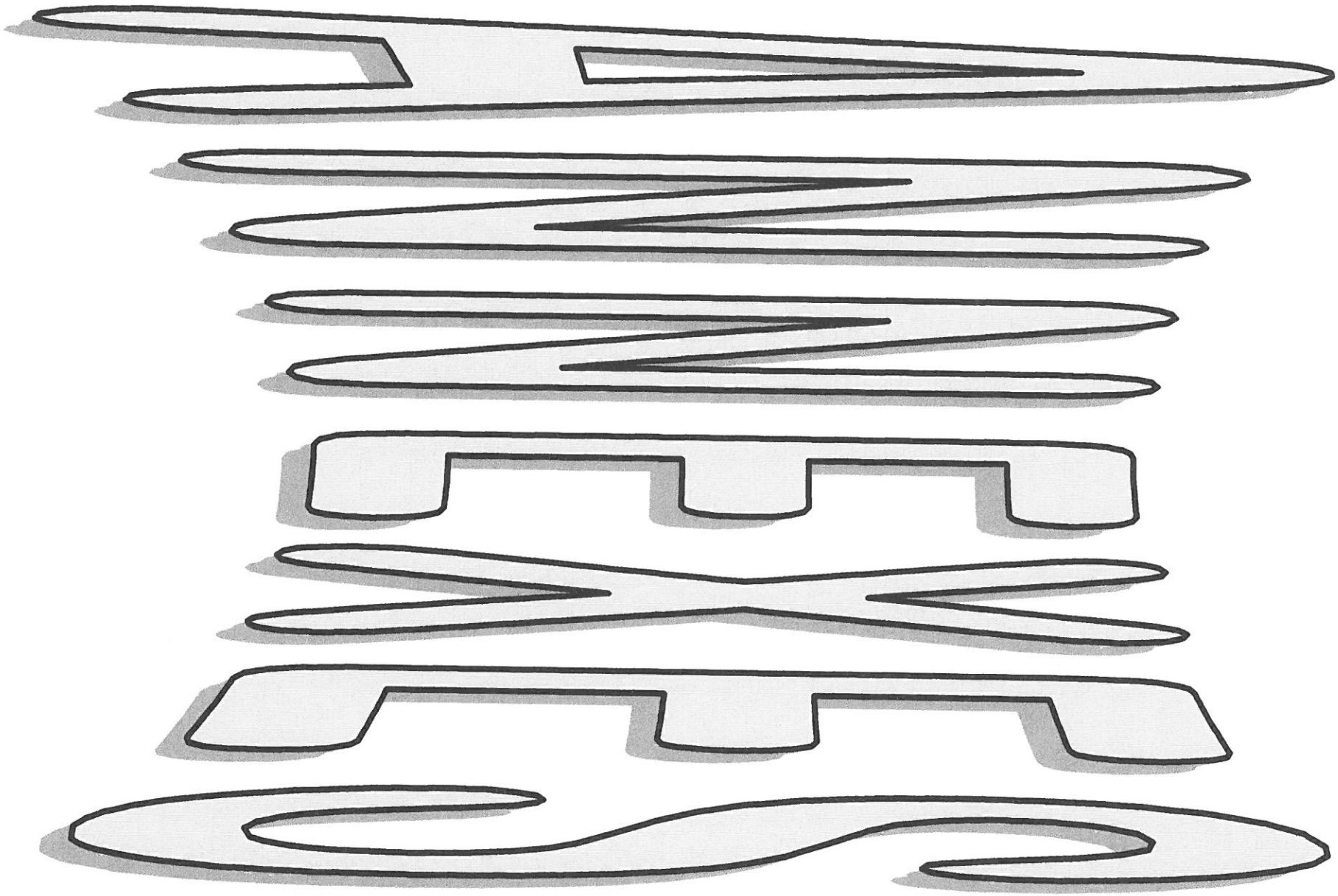
a). Préciser, en justifiant, la loi de probabilité suivie par V et donner ses paramètres.

b). Calculer la probabilité de l'événement "V = 35".

(On donnera l'approximation décimale arrondie à 10^{-2} près de cette probabilité).

c). On remplace la loi de V par une loi normale. Quels sont les paramètres de celle-ci (arrondis à 10^{-2} près) ? On note V' une variable aléatoire qui suit cette loi normale.

En calculant successivement les probabilités de chacun des deux événements "$V' \leq 40,5$" et "$29,5 \leq V' \leq 40,5$", donner des approximations décimales des probabilités de chacun des deux événements "$V \leq 40$" et "$30 \leq V \leq 40$".

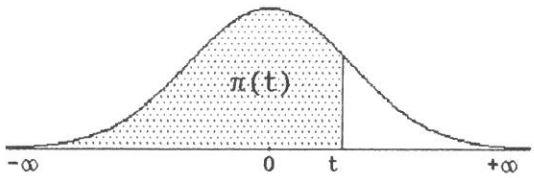


Loi normale

La loi normale centrée réduite a pour densité de probabilité $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$.

Table de la fonction de répartition de la loi normale centrée réduite

$$\pi(t) = \int_{-\infty}^{t} f(x)\, dx$$



t	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986

Grandes valeurs de t

t	3,0	3,1	3,2	3,3	3,4	3,5	3,6	3,8	4,0	4,5
π(t)	0,99865	0,99904	0,99931	0,99952	0,99966	0,99976	0,999841	0,999928	0,999968	0,999997

Fractile de la loi normale centrée réduite

P	0,000	0,001	0,002	0,003	0,004	0,005	0,006	0,007	0,008	0,009	0,010	
0,00	∞	3,0902	2,8782	2,7478	2,6521	2,5758	2,5121	2,4573	2,4089	2,3656	2,3263	0,99
0,01	2,3263	2,2904	2,2571	2,2262	2,1973	2,1701	2,1444	2,1201	2,0969	2,0749	2,0537	0,98
0,02	2,0537	2,0335	2,0141	1,9954	1,9774	1,9600	1,9431	1,9268	1,9110	1,8957	1,8808	0,97
0,03	1,8808	1,8663	1,8522	1,8384	1,8250	1,8119	1,7991	1,7866	1,7744	1,7624	1,7507	0,96
0,04	1,7507	1,7392	1,7279	1,7169	1,7060	1,6954	1,6849	1,6747	1,6646	1,6546	1,6449	0,95
0,05	1,6449	1,6352	1,6258	1,6164	1,6072	1,5982	1,5893	1,5805	1,5718	1,5632	1,5548	0,94
0,06	1,5548	1,5464	1,5382	1,5301	1,5220	1,5141	1,5063	1,4985	1,4909	1,4833	1,4758	0,93
0,07	1,4758	1,4684	1,4611	1,4538	1,4466	1,4395	1,4325	1,4255	1,4187	1,4118	1,4051	0,92
0,08	1,4051	1,3984	1,3917	1,3852	1,3787	1,3722	1,3658	1,3595	1,3532	1,3469	1,3408	0,91
0,09	1,3408	1,3346	1,3285	1,3225	1,3165	1,3106	1,3047	1,2988	1,2930	1,2873	1,2816	0,90
0,10	1,2816	1,2759	1,2702	1,2646	1,2591	1,2536	1,2481	1,2426	1,2372	1,2319	1,2265	0,89
0,11	1,2265	1,2212	1,2160	1,2107	1,2055	1,2004	1,1952	1,1901	1,1850	1,1800	1,1750	0,88
0,12	1,1750	1,1700	1,1650	1,1601	1,1552	1,1503	1,1455	1,1407	1,1359	1,1311	1,1264	0,87
0,13	1,1264	1,1217	1,1170	1,1123	1,1077	1,1031	1,0985	1,0939	1,0893	1,0848	1,0803	0,86
0,14	1,0803	1,0758	1,0714	1,0669	1,0625	1,0581	1,0537	1,0494	1,0450	1,0407	1,0364	0,85
0,15	1,0364	1,0322	1,0279	1,0237	1,0194	1,0152	1,0110	1,0069	1,0027	0,9986	0,9945	0,84
0,16	0,9945	0,9904	0,9863	0,9822	0,9782	0,9741	0,9701	0,9661	0,9621	0,9581	0,9542	0,83
0,17	0,9542	0,9502	0,9463	0,9424	0,9385	0,9346	0,9307	0,9269	0,9230	0,9192	0,9154	0,82
0,18	0,9154	0,9116	0,9078	0,9040	0,9002	0,8965	0,8927	0,8890	0,8853	0,8816	0,8779	0,81
0,19	0,8779	0,8742	0,8705	0,8669	0,8633	0,8596	0,8560	0,8524	0,8488	0,8452	0,8416	0,80
0,20	0,8416	0,8381	0,8345	0,8310	0,8274	0,8239	0,8204	0,8169	0,8134	0,8099	0,8064	0,79
0,21	0,8064	0,8030	0,7995	0,7961	0,7926	0,7892	0,7858	0,7824	0,7790	0,7756	0,7722	0,78
0,22	0,7722	0,7688	0,7655	0,7621	0,7588	0,7554	0,7521	0,7488	0,7454	0,7421	0,7388	0,77
0,23	0,7388	0,7356	0,7323	0,7290	0,7257	0,7225	0,7192	0,7160	0,7128	0,7095	0,7063	0,76
0,24	0,7063	0,7031	0,6999	0,6967	0,6935	0,6903	0,6871	0,6840	0,6808	0,6776	0,6745	0,75
0,25	0,6745	0,6713	0,6682	0,6651	0,6620	0,6588	0,6557	0,6526	0,6495	0,6464	0,6433	0,74
0,26	0,6433	0,6403	0,6372	0,6341	0,6311	0,6280	0,6250	0,6219	0,6189	0,6158	0,6128	0,73
0,27	0,6128	0,6098	0,6068	0,6038	0,6008	0,5978	0,5948	0,5918	0,5888	0,5858	0,5828	0,72
0,28	0,5828	0,5799	0,5769	0,5740	0,5710	0,5681	0,5651	0,5622	0,5592	0,5563	0,5534	0,71
0,29	0,5534	0,5505	0,5476	0,5446	0,5417	0,5388	0,5359	0,5330	0,5302	0,5273	0,5244	0,70
0,30	0,5244	0,5215	0,5187	0,5158	0,5129	0,5101	0,5072	0,5044	0,5015	0,4987	0,4959	0,69
0,31	0,4959	0,4930	0,4902	0,4874	0,4845	0,4817	0,4789	0,4761	0,4733	0,4705	0,4677	0,68
0,32	0,4677	0,4649	0,4621	0,4593	0,4565	0,4538	0,4510	0,4482	0,4454	0,4427	0,4399	0,67
0,33	0,4399	0,4372	0,4344	0,4316	0,4289	0,4261	0,4234	0,4207	0,4179	0,4152	0,4125	0,66
0,34	0,4125	0,4097	0,4070	0,4043	0,4016	0,3989	0,3961	0,3934	0,3907	0,3880	0,3853	0,65
0,35	0,3853	0,3826	0,3799	0,3772	0,3745	0,3719	0,3692	0,3665	0,3638	0,3611	0,3585	0,64
0,36	0,3585	0,3558	0,3531	0,3505	0,3478	0,3451	0,3425	0,3398	0,3372	0,3345	0,3319	0,63
0,37	0,3319	0,3292	0,3266	0,3239	0,3213	0,3186	0,3160	0,3134	0,3107	0,3081	0,3055	0,62
0,38	0,3055	0,3029	0,3002	0,2976	0,2950	0,2924	0,2898	0,2871	0,2845	0,2819	0,2793	0,61
0,39	0,2793	0,2767	0,2741	0,2715	0,2689	0,2663	0,2637	0,2611	0,2585	0,2559	0,2533	0,60
0,40	0,2533	0,2508	0,2482	0,2456	0,2430	0,2404	0,2378	0,2353	0,2327	0,2301	0,2275	0,59
0,41	0,2275	0,2250	0,2224	0,2198	0,2173	0,2147	0,2121	0,2096	0,2070	0,2045	0,2019	0,58
0,42	0,2019	0,1993	0,1968	0,1942	0,1917	0,1891	0,1866	0,1840	0,1815	0,1789	0,1764	0,57
0,43	0,1764	0,1738	0,1713	0,1687	0,1662	0,1637	0,1611	0,1586	0,1560	0,1535	0,1510	0,56
0,44	0,1510	0,1484	0,1459	0,1434	0,1408	0,1383	0,1358	0,1332	0,1307	0,1282	0,1257	0,55
0,45	0,1257	0,1231	0,1206	0,1181	0,1156	0,1130	0,1105	0,1080	0,1055	0,1030	0,1004	0,54
0,46	0,1004	0,0979	0,0954	0,0929	0,0904	0,0878	0,0853	0,0828	0,0803	0,0778	0,0753	0,53
0,47	0,0753	0,0728	0,0702	0,0677	0,0652	0,0627	0,0602	0,0577	0,0552	0,0527	0,0502	0,52
0,48	0,0502	0,0476	0,0451	0,0426	0,0401	0,0376	0,0351	0,0326	0,0301	0,0276	0,0251	0,51
0,49	0,0251	0,0226	0,0201	0,0175	0,0150	0,0125	0,0100	0,0075	0,0050	0,0025	0,0000	0,50
	0,010	0,009	0,008	0,007	0,006	0,005	0,004	0,003	0,002	0,001	0,000	P

Grandes valeurs de u

P	0,9999	0,99999	0,999999	0,9999999	0,99999999	0,999999999
u	3,7190	4,2649	4,7534	5,1993	5,6120	5,9978

Loi de Poisson

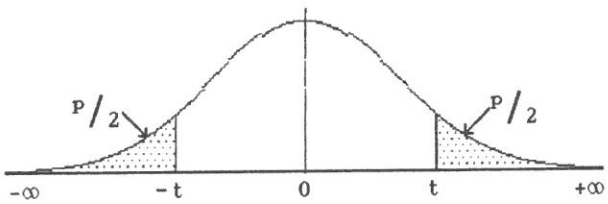
Loi de Poisson de paramètre λ $\quad P[X = k] = \dfrac{e^{-\lambda}\lambda^{k}}{k!}$.

k \ λ	0,2	0,3	0,4	0,5	0,6	0,8
0	0,8187	0,7408	0,6703	0,6065	0,5488	0,4493
1	0,1637	0,2222	0,2681	0,3032	0,3293	0,3595
2	0,0164	0,0333	0,0536	0,0758	0,0988	0,1438
3	0,0011	0,0033	0,0071	0,0126	0,0198	0,0383
4		0,0002	0,0007	0,0015	0,0030	0,0077
5			0,0001	0,0001	0,0004	0,0012
						0,0002

k \ λ	1	1,5	2	3	4	5	6	7	8	9	10
0	0,368	0,223	0,135	0,050	0,018	0,007	0,002	0,001	0,000	0,000	0,000
1	0,368	0,335	0,271	0,149	0,073	0,034	0,015	0,006	0,003	0,001	0,000
2	0,184	0,251	0,271	0,224	0,147	0,084	0,045	0,022	0,011	0,005	0,002
3	0,061	0,126	0,180	0,224	0,195	0,140	0,089	0,052	0,029	0,015	0,008
4	0,015	0,047	0,090	0,168	0,195	0,176	0,134	0,091	0,057	0,034	0,019
5	0,003	0,014	0,036	0,101	0,156	0,176	0,161	0,128	0,092	0,061	0,038
6	0,001	0,004	0,012	0,050	0,104	0,146	0,161	0,149	0,122	0,091	0,063
7	0,000	0,001	0,003	0,022	0,060	0,104	0,138	0,149	0,140	0,117	0,090
8		0,000	0,001	0,008	0,030	0,065	0,103	0,130	0,140	0,132	0,113
9			0,000	0,003	0,013	0,036	0,069	0,101	0,124	0,132	0,125
10				0,001	0,005	0,018	0,041	0,071	0,099	0,119	0,125
11				0,000	0,002	0,008	0,023	0,045	0,072	0,097	0,114
12					0,001	0,003	0,011	0,026	0,048	0,073	0,095
13					0,000	0,001	0,005	0,014	0,030	0,050	0,073
14						0,000	0,002	0,007	0,017	0,032	0,052
15							0,001	0,003	0,009	0,019	0,035
16							0,000	0,001	0,005	0,011	0,022
17								0,000	0,002	0,006	0,013
18									0,001	0,003	0,007
19									0,000	0,001	0,004
20										0,000	0,002
21											0,001
22											0,000

Loi t de Student

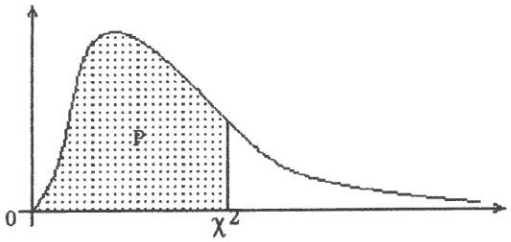
Fractiles de la loi t de Student à ν degrés de liberté.



ν \ P	0,90	0,80	0,70	0,60	0,50	0,40	0,30	0,20	0,10	0,05	0,02	0,01	0,0001
1	0,158	0,325	0,510	0,727	1,000	1,376	1,963	3,078	6,314	12,706	31,821	63,657	636,619
2	0,142	0,289	0,445	0,617	0,816	1,061	1,386	1,886	2,920	4,303	6,695	9,925	31,598
3	0,137	0,277	0,424	0,584	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,841	12,929
4	0,134	0,271	0,414	0,569	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,604	8,610
5	0,132	0,267	0,408	0,559	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,032	6,869
6	0,131	0,265	0,404	0,553	0,718	0,906	1,134	1,440	1,943	2,447	3,143	3,707	5,959
7	0,130	0,263	0,402	0,549	0,711	0,896	1,119	1,415	1,895	2,365	2,998	3,499	5,408
8	0,130	0,262	0,399	0,546	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355	5,041
9	0,129	0,261	0,398	0,543	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250	4,781
10	0,129	0,260	0,397	0,542	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169	4,587
11	0,129	0,260	0,396	0,540	0,697	0,876	1,088	1,363	1,796	2,201	2,718	3,106	4,437
12	0,128	0,259	0,395	0,539	0,695	0,873	1,083	1,356	1,782	2,179	2,681	3,055	4,318
13	0,128	0,259	0,394	0,538	0,694	0,870	1,079	1,350	1,771	2,160	2,650	3,012	4,221
14	0,128	0,258	0,393	0,537	0,692	0,868	1,076	1,345	1,761	2,145	2,624	2,977	4,140
15	0,128	0,258	0,393	0,536	0,691	0,866	1,074	1,341	1,753	2,131	2,602	2,947	4,073
16	0,128	0,258	0,392	0,535	0,690	0,865	1,071	1,337	1,746	2,120	2,583	2,921	4,015
17	0,128	0,257	0,392	0,534	0,689	0,863	1,069	1,333	1,740	2,110	2,567	2,898	3,965
18	0,127	0,257	0,392	0,534	0,688	0,862	1,067	1,330	1,734	2,101	2,552	2,878	3,922
19	0,127	0,257	0,391	0,533	0,688	0,861	1,066	1,328	1,729	2,093	2,539	2,861	3,883
20	0,127	0,257	0,391	0,533	0,687	0,860	1,064	1,325	1,725	2,086	2,528	2,845	3,850
21	0,127	0,257	0,391	0,532	0,686	0,859	1,063	1,323	1,721	2,080	2,518	2,831	3,819
22	0,127	0,256	0,390	0,532	0,686	0,858	1,061	1,321	1,717	2,074	2,508	2,819	3,792
23	0,127	0,256	0,390	0,532	0,685	0,858	1,060	1,319	1,714	2,069	2,500	2,807	3,767
24	0,127	0,256	0,390	0,531	0,685	0,857	1,059	1,318	1,711	2,064	2,492	2,797	3,745
25	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,316	1,708	2,060	2,485	2,787	3,725
26	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,315	1,706	2,056	2,479	2,779	3,707
27	0,127	0,256	0,389	0,531	0,684	0,855	1,057	1,314	1,703	2,052	2,473	2,771	3,690
28	0,127	0,256	0,389	0,530	0,683	0,855	1,056	1,313	1,701	2,048	2,467	2,763	3,674
29	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,311	1,699	2,045	2,462	2,756	3,659
30	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,310	1,697	2,042	2,457	2,750	3,646
40	0,126	0,255	0,388	0,529	0,681	0,851	1,050	1,303	1,684	2,021	2,423	2,704	3,551
80	0,126	0,254	0,387	0,527	0,679	0,848	1,046	1,296	1,671	2,000	2,390	2,660	3,460
120	0,126	0,254	0,386	0,526	0,677	0,845	1,041	1,289	1,658	1,980	2,358	2,617	3,373
∞	0,126	0,253	0,385	0,524	0,674	0,842	1,036	1,282	1,645	1,960	2,326	2,576	3,291

Loi du Khi-deux

Fractiles de la loi du X^2 à ν degrés de liberté.



ν \ P	0,001	0,005	0,01	0,025	0,05	0,1	0,5	0,9	0,95	0,975	0,99	0,995	0,999
1	--	--	--	0,001	0,004	0,016	0,455	2,706	3,841	5,024	6,635	7,879	10,828
2	0,002	0,010	0,020	0,051	0,103	0,211	1,386	4,605	5,991	7,378	9,210	10,597	13,816
3	0,024	0,072	0,115	0,216	0,352	0,584	2,366	6,251	7,815	9,348	11,345	12,838	16,266
4	0,091	0,207	0,297	0,484	0,711	1,064	3,357	7,779	9,488	11,143	13,277	14,860	18,467
5	0,210	0,412	0,554	0,831	1,145	1,610	4,351	9,236	11,070	12,832	15,086	16,750	20,515
6	0,381	0,676	0,872	1,237	1,635	2,204	5,348	10,645	12,592	14,449	16,812	18,548	22,458
7	0,598	0,989	1,239	1,690	2,167	2,833	6,346	12,017	14,067	16,013	18,475	20,278	24,322
8	0,857	1,344	1,646	2,180	2,733	3,490	7,344	13,362	15,507	17,535	20,090	21,955	26,125
9	1,153	1,735	2,088	2,700	3,325	4,168	8,343	14,684	16,919	19,023	21,666	23,589	27,877
10	1,479	2,156	2,558	3,247	3,940	4,865	9,342	15,987	18,307	20,483	23,209	25,188	29,588
11	1,834	2,603	3,053	3,816	4,575	5,578	10,341	17,275	19,675	21,920	24,725	26,757	31,264
12	2,214	3,074	3,571	4,404	5,226	6,304	11,340	18,549	21,026	23,336	26,217	28,300	32,909
13	2,617	3,565	4,107	5,009	5,892	7,042	12,340	19,812	22,362	24,736	27,688	29,819	34,528
14	3,041	4,075	4,660	5,629	6,571	7,790	13,339	21,064	23,685	26,119	29,141	31,319	36,123
15	3,483	4,601	5,229	6,262	7,261	8,547	14,339	22,307	24,996	27,488	30,578	32,801	37,697
16	3,942	5,142	5,812	6,908	7,962	9,312	15,338	23,542	26,296	28,845	32,000	34,267	39,252
17	4,416	5,697	6,408	7,564	8,672	10,085	16,338	24,769	27,587	30,191	33,409	35,718	40,790
18	4,905	6,265	7,015	8,231	9,390	10,865	17,338	25,989	28,869	31,526	34,805	37,156	42,312
19	5,407	6,844	7,633	8,907	10,117	11,651	18,338	27,204	30,144	32,852	36,191	38,582	43,820
20	5,921	7,434	8,260	9,591	10,851	12,443	19,337	28,412	31,410	34,170	37,566	39,997	45,315
21	6,447	8,034	8,897	10,283	11,591	13,240	20,337	29,615	32,671	35,479	38,932	41,401	46,797
22	6,983	8,643	9,542	10,982	12,338	13,041	21,337	30,813	33,924	36,781	40,289	42,796	48,268
23	7,529	9,260	10,196	11,688	13,091	14,848	22,337	32,007	35,172	38,076	41,638	44,181	49,728
24	8,085	9,886	10,856	12,401	13,848	15,659	23,337	33,196	36,415	39,364	42,980	45,558	51,179
25	8,649	10,520	11,524	13,120	14,611	16,473	24,337	34,382	37,652	40,646	44,314	46,928	52,620
26	9,222	11,160	12,198	13,844	15,379	17,292	25,336	35,563	38,885	41,923	45,642	48,290	54,052
27	9,803	11,808	12,879	14,573	16,151	18,114	26,336	36,741	40,113	43,194	46,963	49,645	55,476
28	10,391	12,461	13,565	15,308	16,928	18,939	27,336	37,916	41,337	44,461	48,278	50,993	56,892
29	10,986	13,121	14,256	16,047	17,708	19,768	28,336	39,087	42,557	45,722	49,588	52,336	58,302
30	11,588	13,787	14,953	16,791	18,493	20,599	29,336	40,256	43,773	46,979	50,892	53,672	59,703

Programmes des S.T.S.

B.O. du 25 mai 1989

CALCUL DES PROBABILITES 1

Il s'agit *d'une initiation aux phénomènes aléatoires où toute ambition théorique et toute technicité sont exclues.* L'objectif est que les élèves sachent traiter quelques problèmes simples concernant des variables aléatoires dont la loi figure au programme et utiliser les tables de ces lois. Les sciences et techniques industrielles et économiques fournissent un large éventail de tels problèmes, et on évitera les situations artificielles.

a) Probabilités sur les ensembles finis : vocabulaire des événements, probabilité. Probabilité conditionnelle, événements indépendants. Cas équiprobable. Arrangements, combinaisons.	L'ensemble des événements sera pris égal à l'ensemble de toutes les parties de Ω. Aucune difficulté ne sera soulevée sur l'extension aux ensembles infinis. Ces notions sont introduites lors de l'étude de quelques situations simples de dénombrement. On pourra étudier les algorithmes associés (calcul de n! découlant de la relation $n! = (n-1)!\, n$, ...)
b) Variables aléatoires à valeurs réelles : loi de probabilité, fonction de répartition. Espérance mathématique, variance, écart type. Loi binomiale, loi de Poisson, loi normale.	Aucune difficulté théorique ne doit être soulevée sur les variables aléatoires. A propos de la loi normale, on sera amené à utiliser les notations $\int_a^{+\infty} f(t)\,dt$, $\int_{-\infty}^{b} f(t)\,dt$, $\int_{-\infty}^{+\infty} f(t)\,dt$, mais aucune connaissance sur les intégrales impropres n'est exigible des élèves.

Travaux pratiques

1. Emploi de dénombrements pour le calcul de probabilités.	On se limitera à des exemples simples.
2. Exemples d'études de situations de probabilités faisant intervenir des variables aléatoires suivant une loi binomiale, de Poisson ou normale.	Les élèves doivent savoir reconnaître qu'un phénomène suit une loi binomiale et remplacer éventuellement celle-ci par une loi de Poisson ou une loi normale.

CALCUL DES PROBABILITES 2

Il s'agit *d'une initiation aux phénomènes aléatoires où toute ambition théorique et toute technicité sont exclues.* L'objectif est que les élèves sachent traiter quelques problèmes simples concernant des variables aléatoires dont la loi figure au programme et utiliser les tables de ces lois. Les sciences et techniques industrielles et économiques fournissent un large éventail de tels problèmes, et on évitera les situations artificielles

a) Probabilités sur les ensembles finis : vocabulaire des événements, probabilité. Probabilité conditionnelle, événements indépendants. Cas équiprobable. Arrangements, combinaisons.	L'ensemble des événements sera pris égal à l'ensemble de toutes les parties de Ω. Aucune difficulté ne sera soulevée sur l'extension aux ensembles infinis. Ces notions sont introduites lors de l'étude de quelques situations simples de dénombrement. On pourra étudier les algorithmes associés (calcul de n! découlant de la relation $n! = (n-1)!\, n$, ...)
b) Variables aléatoires à valeurs réelles : loi de probabilité, fonction de répartition. Espérance mathématique, variance, écart type. Loi binomiale, loi de Poisson, loi normale.	Aucune difficulté théorique ne doit être soulevée sur les variables aléatoires. A propos de la loi normale, on sera amené à utiliser les notations $\int_a^{+\infty} f(t)\,dt$, $\int_{-\infty}^{b} f(t)\,dt$, $\int_{-\infty}^{+\infty} f(t)\,dt$, mais aucune connaissance sur les intégrales impropres n'est exigible des élèves.

Somme de deux variables aléatoires, espérance de la somme ; indépendance de deux variables aléatoires, variance de la somme de deux variables aléatoires indépendantes	On donnera aussi les résultats pour la différence de deux variables aléatoires.
c) Enoncés de la loi faible des grands nombres et du théorème de la limite centrée ; interprétation statistique : distribution d'échantillonnage des moyennes, des pourcentages.	Aucune difficulté théorique ne doit être soulevée à propos de la convergence d'une suite de variables aléatoires

Travaux pratiques

1. Emploi de dénombrements pour le calcul de probabilités.	On se limitera à des exemples simples.
2. Exemples d'études de situations de probabilités faisant intervenir des variables aléatoires suivant une loi binomiale, de Poisson ou normale.	Les élèves doivent savoir reconnaître qu'un phénomène suit une loi binomiale et remplacer éventuellement celle-ci par une loi de Poisson ou une loi normale. En liaison avec l'enseignement technologique on pourra être amené à utiliser d'autres lois, notamment la loi log-normale, mais aucune connaissance n'est exigible à ce sujet en mathématiques.

STATISTIQUE INFERENTIELLE 1

L'objectif essentiel de ce chapitre est d'initier les élèves à l'utilisation de méthodes statistiques pour estimer un paramètre. A l'aide des modèles théoriques que le calcul des probabilités permet de dégager, les notions de statistique inférentielle figurant au programme ont pour but la prise d'une décision concernant l'acceptation ou le refus d'une hypothèse.

On considérera uniquement de grands échantillons et on se limitera aux situations où les seules lois continues à envisager sont des lois normales.

a) Estimation ponctuelle d'un paramètre : moyenne, pourcentage, variance, écart type.	Toute étude des qualités (biais, convergence, ..) d'un estimateur est hors programme.
b) Estimation par intervalle de confiance d'une moyenne ou d'un pourcentage.	On entraînera les élèves à utiliser une terminologie correcte en évitant de confondre la probabilité pour que la valeur d'un paramètre appartienne à un intervalle de confiance avec un pourcentage portant sur la population.
c) Test de validité d'hypothèse relatif à une moyenne ou un pourcentage. Application à la comparaison de deux populations.	La comparaison de deux populations est à lier en particulier à la recherche d'une variation significative d'un paramètre.

Travaux pratiques

1. Exemples d'estimation d'une moyenne ou d'un pourcentage par un intervalle de confiance.	On évitera les situations artificielles en privilégiant les exemples issus de la vie économique et sociale (enquêtes, sondages, ...)
2. Exemples de construction et d'utilisation d'un test de validité d'hypothèse relatif à une moyenne ou à un pourcentage.	

STATISTIQUE INFERENTIELLE 2

L'objectif essentiel de ce chapitre est d'initier les élèves à l'utilisation de méthodes statistiques pour contrôler la qualité d'une fabrication. A l'aide des modèles théoriques que le calcul des probabilités permet de dégager, les notions de statistique inférentielle figurant au programme ont pour but la prise d'une décision concernant la conformité d'une production à un cahier des charges soit en cours de fabrication, soit lors de la réception de la marchandise, et l'estimation de la durée de vie d'un équipement. Dans les problèmes d'estimation par intervalle de confiance les *exigences en mathématiques ne concernent que les situations où les seules lois continues à envisager sont des lois normales et où l'on utilise de grands échantillons*. En liaison avec l'enseignement technologique on pourra être amené, d'une part, à considérer de petits échantillons et à utiliser la loi de Student et, d'autre part, à étudier graphiquement une distribution expérimentale (droite de Henry), mais ces questions ne figurent pas au programme de mathématiques.

a) Estimation ponctuelle d'un paramètre : moyenne, pourcentage, variance, écart type.	Toute étude des qualités (biais, convergence, ..) d'un estimateur est hors programme.
b) Estimation par intervalle de confiance d'une moyenne ou d'une fréquence (d'un pourcentage).	On entraînera les élèves à utiliser une terminologie correcte en évitant de confondre la probabilité pour que la valeur d'un paramètre appartienne à un intervalle de confiance avec un pourcentage portant sur la population.
c) Test de validité d'hypothèse relatif à une moyenne ou une fréquence (d'un pourcentage). Application au contrôle de la qualité d'une fabrication et à la comparaison de deux populations.	La comparaison de deux populations est à lier en particulier à la recherche d'une amélioration de la qualité d'une fabrication.
d) Fiabilité d'un système ou d'un élément à un instant donné dans le cas d'un taux d'avarie constant.	Dans cette brève étude de fiabilité on introduira la loi exponentielle. La loi de Weibull est hors programme. L'objectif est ici de savoir effectuer un calcul de durée de vie moyenne et une estimation ponctuelle de cette durée par un essai.

Travaux pratiques

1. Exemples d'estimation d'une moyenne ou d'une fréquence par un intervalle de confiance.	On évitera les situations artificielles et on effectuera le lien entre intervalle de confiance et tolérance.
2. Construction et utilisation d'un test de validité d'hypothèse relatif à une moyenne ou à une fréquence.	On considérera notamment les cas suivants : - contrôle par mesures : test sur la moyenne ; - contrôle par "attributs" : test sur le pourcentage d'éléments "défectueux"
3. Exemples d'étude de fiabilité à l'aide de la loi exponentielle.	On observera que l'utilisation de papier semi-logarithmique permet en particulier de déterminer graphiquement l'espérance mathématique (MTBF).

STATISTIQUE INFERENTIELLE 3

A l'aide des modèles théoriques que le calcul des probabilités permet de dégager, les notions de statistique inférentielle figurant au programme ont pour but d'étudier la fiabilité d'un équipement et la qualité de son fonctionnement.

Dans les problèmes d'estimation par intervalle de confiance les *exigences en mathématiques ne concernent que les situations où les seules lois continues à envisager sont des lois normales et où l'on utilise de grands échantillons*. En liaison avec l'enseignement technologique on pourra être amené, d'une part, à considérer de petits échantillons et à utiliser la loi de Student et, d'autre part, à étudier graphiquement une distribution expérimentale (droite de Henry), mais ces questions ne figurent pas au programme de mathématiques.

a) Estimation ponctuelle d'un paramètre : moyenne, pourcentage, variance, écart type.	Toute étude des qualités (biais, convergence, ..) d'un estimateur est hors programme.

b) Estimation par intervalle de confiance d'une moyenne ou d'une fréquence (d'un pourcentage).	On entraînera les élèves à utiliser une terminologie correcte en évitant de confondre la probabilité pour que la valeur d'un paramètre appartienne à un intervalle de confiance avec un pourcentage portant sur la population.
c) Test de validité d'hypothèse relatif à une moyenne ou une fréquence (un pourcentage). Application au contrôle de la qualité d'une fabrication et à la comparaison de deux populations.	La comparaison de deux populations est à lier en particulier à la recherche d'une amélioration de la qualité d'une fabrication.
d) Fiabilité d'un élément ou d'un système : fonction de fiabilité R(t), fonction de défaillance F(t), taux d'avarie $\lambda(t)$. Loi exponentielle dans le cas d'un taux d'avarie constant : densité de probabilité, espérance mathématique (MTBF), écart type. Loi de Weibull : taux d'avarie, fonction de fiabilité, fonction de défaillance, densité de probabilité, espérance mathématique (MTBF), écart type, influence des paramètres γ, β, η.	Cette étude est à mener en liaison étroite avec l'enseignement de la maintenance qui fournit un large éventail de situations où ces notions sont utilisées. On illustrera graphiquement ces définitions.

Travaux pratiques

1. Exemples d'estimation d'une moyenne ou d'une fréquence par un intervalle de confiance.	On évitera les situations artificielles et on effectuera le lien entre intervalle de confiance et tolérance.
2. Exemples de construction et d'utilisation d'un test de validité d'hypothèse relatif à une moyenne ou à une fréquence.	On considérera notamment les cas suivants : - contrôle par mesures : test sur la moyenne, - contrôle par "attributs" : test sur le pourcentage d'éléments "défectueux"
3. Exemples d'étude de fiabilité à l'aide de la loi exponentielle ou de la loi de Weibull (fiabilité d'un composant, fiabilité d'un système en fonction de celle de ses composants).	On observera que l'utilisation de papier semi-logarithmique permet en particulier de déterminer graphiquement l'espérance mathématique (MTBF). On utilisera le papier de Weibull pour déterminer graphiquement les paramètres γ, β, η, puis des tables pour déterminer l'espérance mathématique (MTBF) et l'écart type. Toute recherche de modélisation d'une distribution par une loi de Weibull, en particulier l'utilisation du test de Kolmogorov-Smirnov, est hors programme.

Un Formulaire T.S. expérimenté en classe (non officiel)

Modules : 1 & 2 Calcul des probabilités ◆ 1 & 2 Statistiques inférentielles ◆ Statistiques descriptives

I. STATISTIQUE

Moyenne, variance, écart type

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n} x_i \quad ; \quad V(x)=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2=\frac{1}{n}\sum_{i=1}^{n}x_i^2-(\bar{x})^2$$

Dans le cas d'un regroupement en classes :

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{p} n_i x_i \; ; \; V(x)=\frac{1}{n}\sum_{i=1}^{p} n_i(x_i-\bar{x})^2=\frac{1}{n}\sum_{i=1}^{p} n_i x_i^2-(\bar{x})^2$$

Dans tous les cas : $\sigma(x)=\sqrt{V(x)}$

Droites de régression

$$\mathrm{Cov}(x,y)=\sigma_{xy}=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})=\frac{1}{n}\sum_{i=1}^{n}(x_i y_i)-\bar{x}\,\bar{y}$$

$y=ax+b$, où $a=\dfrac{\mathrm{Cov}(x,y)}{V(x)}$; $x=a'y+b'$, où $a'=\dfrac{\mathrm{Cov}(x,y)}{V(y)}$

Coefficient de corrélation linéaire $r=\dfrac{\mathrm{Cov}(x,y)}{\sigma_x\sigma_y}$

II. COMBINATOIRE - DÉNOMBREMENTS

$\mathrm{Card}(A\cup B)=\mathrm{Card}\,A+\mathrm{Card}\,B-\mathrm{Card}(A\cap B)$

$\mathrm{Card}(A\times B)=\mathrm{Card}\,A\times\mathrm{Card}\,B$

Soit E un ensemble de n éléments

Nombre de permutations de E : $n!=1\times2\times3\times\ldots\times n$; $0!=1$

Nombre d'arrangements de p éléments de E :

$$A_n^p=n(n-1)\ldots(n-p+1)$$

Nombre de sous-ensembles de p éléments de E :

$$C_n^p=\binom{n}{p}=\frac{A_n^p}{p!}=\frac{n(n-1)\ \ldots\ (n-p+1)}{p!}=\frac{n!}{p!\,(n-p)!}$$

$$C_n^p=C_n^{n-p} \quad ; \quad C_{n+1}^{p+1}=C_n^p+C_n^{p+1}$$

$$(a+b)^n=a^n+C_n^1a^{n-1}b+\ldots+C_n^k a^{n-k}b^k+\ldots+b^n$$

III. PROBABILITÉS

Si A et B sont incompatibles : $P(A\cup B)=P(A)+P(B)$

Dans le cas général : $P(A\cup B)=P(A)+P(B)-P(A\cap B)$

$P(\bar{A})=1-P(A)$; $P(\Omega)=1$; $P(\varnothing)=0$

Si $A_1,\ldots,A_n$ forment une partition de A, $P(A)=\sum_{i=1}^{n}P(A_i)$

Dans le cas équiprobable : $P(A)=\dfrac{\mathrm{Card}\,A}{\mathrm{Card}\,\Omega}$

Probabilité conditionnelle de A sachant que B est réalisé

$P(A\cap B)=P(A\,|\,B)\,P(B)$; $P(A\,|\,B)$ se note aussi $P_B(A)$

Cas où A et B sont indépendants : $P(A\cap B)=P(A)\,P(B)$

Formule des probabilités totales

Si les événements $B_1, B_2, \ldots, B_n$ forment une partition de Ω,

alors $P(A)=P(A\cap B_1)+P(A\cap B_2)+\ldots+P(A\cap B_n)$

Variable aléatoire

Fonction de répartition : $F(x)=P(X\leq x)$

Espérance mathématique : $E(X)=\sum_{i=1}^{n}p_i x_i$

Variance : $V(X)=\sum_{i=1}^{n}p_i(x_i-E(X))^2=\sum_{i=1}^{n}p_i x_i^2-(E(X))^2$

Écart type : $\sigma_x=\sqrt{V(X)}$

Loi binomiale : $X(\Omega)=\{0,1,\ldots,n\}$; $P(X=k)=C_n^k p^k(1-p)^{n-k}$

$E(X)=np$; $V(X)=np(1-p)$

Loi de Poisson : $X(\Omega)=\mathbb{N}$; $P(X=k)=\dfrac{\lambda^k e^{-\lambda}}{k!}$; $E(X)=\lambda$; $V(X)=\lambda$

Si X_1 suit la loi $\mathcal{N}(m_1,\sigma_1)$, X_2 suit la loi $\mathcal{N}(m_2,\sigma_2)$, et sont indépendantes, alors X_1+X_2 suit la loi $\mathcal{N}(m_1+m_2,\sqrt{\sigma_1^2+\sigma_2^2})$

Échantillonnage

◆ La variable aléatoire $\bar{X}$ qui, à tout échantillon aléatoire **non exhaustif** de taille n prélevé dans une population de moyenne m et d'écart type σ, associe la moyenne de cet échantillon, pour n assez grand, suit approximativement la loi $\mathcal{N}\left(m,\dfrac{\sigma}{\sqrt{n}}\right)$.

◆ La variable aléatoire F qui, à tout échantillon aléatoire **non exhaustif** de taille n prélevé dans une population présentant un pourcentage p d'individus possédant une certaine propriété, associe le pourcentage des éléments de cet échantillon qui possèdent cette propriété, pour n assez grand,

suit approximativement la loi $\mathcal{N}\left(p,\sqrt{\dfrac{p(1-p)}{n}}\right)$.

◆ Dans le cas d'échantillons **exhaustifs** les écarts types de $\bar{X}$ et de F sont multipliés par le facteur d'exhaustivité $\sqrt{\dfrac{N-n}{N-1}}$.

IV. STATISTIQUE INFÉRENTIELLE

Estimation ponctuelle

◆ Une estimation de la moyenne m d'une population, est donnée par la moyenne $\bar{x}$ d'un échantillon aléatoire non exhaustif prélevé dans cette population.

◆ Une estimation du pourcentage p d'individus possédant une certaine propriété dans une population, est donnée par le pourcentage f d'individus possédant cette propriété dans un échantillon aléatoire non exhaustif prélevé dans cette population.

◆ Une estimation de l'écart type σ d'une population, est donnée par $\sqrt{\dfrac{n}{n-1}}\,s$, où n est la taille et s l'écart type d'un échantillon aléatoire non exhaustif prélevé dans cette population.

Estimation par intervalle de confiance

◆ Un intervalle de confiance de la moyenne m d'une population d'écart type σ, avec le coefficient de confiance $2\Pi(t)-1$, est

$\left[\bar{x}-t\dfrac{\sigma}{\sqrt{n}}\ ,\ \bar{x}+t\dfrac{\sigma}{\sqrt{n}}\right]$, où n est la taille et $\bar{x}$ la moyenne d'un échantillon aléatoire non exhaustif prélevé dans cette population.

◆ Un intervalle de confiance d'un pourcentage p d'individus possédant une certaine propriété dans une population, avec le coefficient de confiance $2\Pi(t)-1$, est

$$\left[f-t\sqrt{\frac{f(1-f)}{n-1}}\ ,\ f+t\sqrt{\frac{f(1-f)}{n-1}}\right]$$

où n est la taille d'un échantillon aléatoire non exhaustif prélevé dans la population, et f le pourcentage d'individus possédant cette propriété dans cet échantillon.

◆ Dans le cas d'échantillons **exhaustifs** les intervalles de confiances décrits ci dessus sont respectivement :

$$\left[\bar{x}-t\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}\ ,\ \bar{x}+t\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}\right]$$

$$\left[f-t\sqrt{\frac{f(1-f)}{n-1}}\sqrt{\frac{N-n}{N-1}}\ ,\ f+t\sqrt{\frac{f(1-f)}{n-1}}\sqrt{\frac{N-n}{N-1}}\right]$$

Tests d'hypothèse

◆ Choix : - d'une hypothèse nulle H_0 ($m=m_0$ ou $p=p_0$)

- d'une alternative H_1 ($m\neq m_0$ ou $p\neq p_0$) [test bilatéral]

($m>m_0$, $m<m_0$, $p>p_0$, $p<p_0$) [test unilatéral]

◆ Recherche de la région d'acceptation ou de la région critique sous H_0 au seuil de risque α

◆ Énoncé de la règle de décision

◆ Application du test avec échantillon(s)

Bibliographie

Probabilité et inférence statistique	N. ABBOUD et J.F. AUDROING Nathan Supérieur Economie
Statistique et probabilités pour aujourd'hui	Irwing ADLER O.C.D.L. Paris 1963
Mathématiques statistiques-probabilités Collection Demengel Formations supérieures technologiques et tertiaires	P. BENICHOU, R. BENICHOU, N. BOY et J.P. POUGET Dunod
Mathématiques pour les sciences de la nature et de la vie	Françoise et Jean-Paul BERTRANDIAS Presses Universitaires de Grenoble 1990
Statistique et probabilités	Bernard BIGOT et Bernard VERLANT Foucher 1993
Dossiers pour la formation continue Estimation - Tests	Michel CHAVIGNY C.A.F.O.C. de Besançon
Les probabilités à l'école	Maurice GLAYMANN et Tamas VARGA CEDIC 1975
Au hasard : la chance, la science et le monde	Ivar EKELAND Seuil 1991
Les certitudes du hasard	Arthur ENGEL Aléas 1990
Probabilités fortuites Exercices et problèmes ordinaires avec solutions et rappels de cours	Gérard FRUGIER Ellipses 1993
Exercices ordinaires de probabilités	Gérard FRUGIER Ellipses 1992
Statistique appliquée à la gestion	Vincent GIARD Economica
Cours de probabilités et statistique 2ème édition	J. GUEGAND, C. LEBOEUF et J.L. ROQUE Ellipses 1987
Cours de probabilités - Licence	Michel HENRY C.T.U. de Besançon
L'enseignement des statistiques et des probabilités dans les S.T.S.	I.R.E.M. de Besançon 1990
Fiabilité (brochure n° 48)	I.R.E.M. de Paris Nord 1992 Groupe Inter-I.R.E.M. Lycées Techniques
Des textes avec corrigés des B.T.S. rénovés des sessions de 1990 et 1991 (brochure n° 58)	I.R.E.M. de Paris Nord Groupe Inter-I.R.E.M. Lycées Techniques

Actes de l'université d'été de statistique	I.R.EM. de Rouen 1992
Statistique - Information - Estimation -Tests Economie Module	Pascal KAUFFMANN Dunod 1994
Attention statistique	Joseph KLATZMANN Cahiers libres / La découverte
Statistique et probabilités Précis de cours accompagné de 120 exercices corrigés	Michel LAVIEVILLE Dunod Université 1990
Techniques statistiques	Georges PARREINS Dunod Technique 1974
Théorie et méthodes de la statistique	G. SAPORTA Technip
La statistique, outil de la qualité	P. SOUVAY Afnor Gestion 1986
Statistique - Cours et problèmes	Murray R. SPIEGEL Série Schaum 1993
Statistique Exercices d'application 4ème édition	T.H. et R.J. WONNACOTT Economica 1991
Statistique - Economie - Gestion - Sciences - Médecine 3ème édition	T.H. et R.J. WONNACOTT Economica 1988
Annales corrigées B.T.S. tertiaires	Foucher

TITRE : PROBABILITES ET STATISTIQUES - STATISTIQUES INFERENTIELLES (B.T.S.)

AUTEUR : B. BIGOT, B. CHAPUT, J.C. DUPERRET, J.C. DANIEL

NIVEAU : Lycée : sections de B.T.S.

EDITEUR : IREM de Reims

DATE : Octobre 1996

MOTS-CLE : spécialité PROBABILITES, STATISTIQUES, STATISTIQUES INFERENTIELLES

autres
- Estimation ponctuelle et par intervalle
- Loi (de probabilité)
- Statistiques et statistiques inférentielles
- Test d'hypothèses
- Variable aléatoire

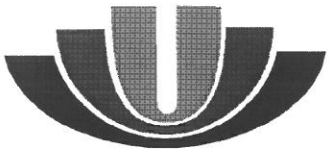
RESUME : Cette brochure a été réalisée à partir des documents élaborés pour l'animation de stages de formation des enseignants de mathématiques des sections de Techniciens Supérieurs.

Elle est le fruit des recherches du groupe B.T.S. de l'IREM de Reims et est avant tout destinée aux enseignants.

La première partie pose le problème de la modélisation (passage des statistiques aux probabilités), la seconde aborde la notion de loi de probabilité et débouche sur l'échantillonnage en vue des statistiques inférentielles traitées dans la troisième partie.

ISBN 2-910076-10-5

FORMAT	NOMBRE DE PAGES	PRIX	IREM numéro
A4	174	75 F + 16 F (frais d'envoi)	Re37



Fractile de la loi normale centrée réduite

P	0,000	0,001	0,002	0,003	0,004	0,005	0,006	0,007	0,008	0,009	0,010	
0,00	∞	3,0902	2,8782	2,7478	2,6521	2,5758	2,5121	2,4573	2,4089	2,3656	2,3263	0,99
0,01	2,3263	2,2904	2,2571	2,2262	2,1973	2,1701	2,1444	2,1201	2,0969	2,0749	2,0537	0,98
0,02	2,0537	2,0335	2,0141	1,9954	1,9774	1,9600	1,9431	1,9268	1,9110	1,8957	1,8808	0,97
0,03	1,8808	1,8663	1,8522	1,8384	1,8250	1,8119	1,7991	1,7866	1,7744	1,7624	1,7507	0,96
0,04	1,7507	1,7392	1,7279	1,7169	1,7060	1,6954	1,6849	1,6747	1,6646	1,6546	1,6449	0,95
0,05	1,6449	1,6352	1,6258	1,6164	1,6072	1,5982	1,5893	1,5805	1,5718	1,5632	1,5548	0,94
0,06	1,5548	1,5464	1,5382	1,5301	1,5220	1,5141	1,5063	1,4985	1,4909	1,4833	1,4758	0,93
0,07	1,4758	1,4684	1,4611	1,4538	1,4466	1,4395	1,4325	1,4255	1,4187	1,4118	1,4051	0,92
0,08	1,4051	1,3984	1,3917	1,3852	1,3787	1,3722	1,3658	1,3595	1,3532	1,3469	1,3408	0,91
0,09	1,3408	1,3346	1,3285	1,3225	1,3165	1,3106	1,3047	1,2988	1,2930	1,2873	1,2816	0,90
0,10	1,2816	1,2759	1,2702	1,2646	1,2591	1,2536	1,2481	1,2426	1,2372	1,2319	1,2265	0,89
0,11	1,2265	1,2212	1,2160	1,2107	1,2055	1,2004	1,1952	1,1901	1,1850	1,1800	1,1750	0,88
0,12	1,1750	1,1700	1,1650	1,1601	1,1552	1,1503	1,1455	1,1407	1,1359	1,1311	1,1264	0,87
0,13	1,1264	1,1217	1,1170	1,1123	1,1077	1,1031	1,0985	1,0939	1,0893	1,0848	1,0803	0,86
0,14	1,0803	1,0758	1,0714	1,0669	1,0625	1,0581	1,0537	1,0494	1,0450	1,0407	1,0364	0,85
0,15	1,0364	1,0322	1,0279	1,0237	1,0194	1,0152	1,0110	1,0069	1,0027	0,9986	0,9945	0,84
0,16	0,9945	0,9904	0,9863	0,9822	0,9782	0,9741	0,9701	0,9661	0,9621	0,9581	0,9542	0,83
0,17	0,9542	0,9502	0,9463	0,9424	0,9385	0,9346	0,9307	0,9269	0,9230	0,9192	0,9154	0,82
0,18	0,9154	0,9116	0,9078	0,9040	0,9002	0,8965	0,8927	0,8890	0,8853	0,8816	0,8779	0,81
0,19	0,8779	0,8742	0,8705	0,8669	0,8633	0,8596	0,8560	0,8524	0,8488	0,8452	0,8416	0,80
0,20	0,8416	0,8381	0,8345	0,8310	0,8274	0,8239	0,8204	0,8169	0,8134	0,8099	0,8064	0,79
0,21	0,8064	0,8030	0,7995	0,7961	0,7926	0,7892	0,7858	0,7824	0,7790	0,7756	0,7722	0,78
0,22	0,7722	0,7688	0,7655	0,7621	0,7588	0,7554	0,7521	0,7488	0,7454	0,7421	0,7388	0,77
0,23	0,7388	0,7356	0,7323	0,7290	0,7257	0,7225	0,7192	0,7160	0,7128	0,7095	0,7063	0,76
0,24	0,7063	0,7031	0,6999	0,6967	0,6935	0,6903	0,6871	0,6840	0,6808	0,6776	0,6745	0,75
0,25	0,6745	0,6713	0,6682	0,6651	0,6620	0,6588	0,6557	0,6526	0,6495	0,6464	0,6433	0,74
0,26	0,6433	0,6403	0,6372	0,6341	0,6311	0,6280	0,6250	0,6219	0,6189	0,6158	0,6128	0,73
0,27	0,6128	0,6098	0,6068	0,6038	0,6008	0,5978	0,5948	0,5918	0,5888	0,5858	0,5828	0,72
0,28	0,5828	0,5799	0,5769	0,5740	0,5710	0,5681	0,5651	0,5622	0,5592	0,5563	0,5534	0,71
0,29	0,5534	0,5505	0,5476	0,5446	0,5417	0,5388	0,5359	0,5330	0,5302	0,5273	0,5244	0,70
0,30	0,5244	0,5215	0,5187	0,5158	0,5129	0,5101	0,5072	0,5044	0,5015	0,4987	0,4959	0,69
0,31	0,4959	0,4930	0,4902	0,4874	0,4845	0,4817	0,4789	0,4761	0,4733	0,4705	0,4677	0,68
0,32	0,4677	0,4649	0,4621	0,4593	0,4565	0,4538	0,4510	0,4482	0,4454	0,4427	0,4399	0,67
0,33	0,4399	0,4372	0,4344	0,4316	0,4289	0,4261	0,4234	0,4207	0,4179	0,4152	0,4125	0,66
0,34	0,4125	0,4097	0,4070	0,4043	0,4016	0,3989	0,3961	0,3934	0,3907	0,3880	0,3853	0,65
0,35	0,3853	0,3826	0,3799	0,3772	0,3745	0,3719	0,3692	0,3665	0,3638	0,3611	0,3585	0,64
0,36	0,3585	0,3558	0,3531	0,3505	0,3478	0,3451	0,3425	0,3398	0,3372	0,3345	0,3319	0,63
0,37	0,3319	0,3292	0,3266	0,3239	0,3213	0,3186	0,3160	0,3134	0,3107	0,3081	0,3055	0,62
0,38	0,3055	0,3029	0,3002	0,2976	0,2950	0,2924	0,2898	0,2871	0,2845	0,2819	0,2793	0,61
0,39	0,2793	0,2767	0,2741	0,2715	0,2689	0,2663	0,2637	0,2611	0,2585	0,2559	0,2533	0,60
0,40	0,2533	0,2508	0,2482	0,2456	0,2430	0,2404	0,2378	0,2353	0,2327	0,2301	0,2275	0,59
0,41	0,2275	0,2250	0,2224	0,2198	0,2173	0,2147	0,2121	0,2096	0,2070	0,2045	0,2019	0,58
0,42	0,2019	0,1993	0,1968	0,1942	0,1917	0,1891	0,1866	0,1840	0,1815	0,1789	0,1764	0,57
0,43	0,1764	0,1738	0,1713	0,1687	0,1662	0,1637	0,1611	0,1586	0,1560	0,1535	0,1510	0,56
0,44	0,1510	0,1484	0,1459	0,1434	0,1408	0,1383	0,1358	0,1332	0,1307	0,1282	0,1257	0,55
0,45	0,1257	0,1231	0,1206	0,1181	0,1156	0,1130	0,1105	0,1080	0,1055	0,1030	0,1004	0,54
0,46	0,1004	0,0979	0,0954	0,0929	0,0904	0,0878	0,0853	0,0828	0,0803	0,0778	0,0753	0,53
0,47	0,0753	0,0728	0,0702	0,0677	0,0652	0,0627	0,0602	0,0577	0,0552	0,0527	0,0502	0,52
0,48	0,0502	0,0476	0,0451	0,0426	0,0401	0,0376	0,0351	0,0326	0,0301	0,0276	0,0251	0,51
0,49	0,0251	0,0226	0,0201	0,0175	0,0150	0,0125	0,0100	0,0075	0,0050	0,0025	0,0000	0,50
	0,010	0,009	0,008	0,007	0,006	0,005	0,004	0,003	0,002	0,001	0,000	P

Grandes valeurs de u

P	0,9999	0,99999	0,999999	0,9999999	0,99999999	0,999999999
u	3,7190	4,2649	4,7534	5,1993	5,6120	5,9978

Un Formulaire T.S. expérimenté en classe (non officiel)

Modules : 1 & 2 Calcul des probabilités ◆ 1 & 2 Statistiques inférentielles ◆ Statistiques descriptives

I. STATISTIQUE

Moyenne, variance, écart type

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i \quad ; \quad V(x)=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2=\frac{1}{n}\sum_{i=1}^{n}x_i^2-(\bar{x})^2$$

Dans le cas d'un regroupement en classes :

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{p}n_i x_i \; ; \; V(x)=\frac{1}{n}\sum_{i=1}^{p}n_i(x_i-\bar{x})^2=\frac{1}{n}\sum_{i=1}^{p}n_i x_i^2-(\bar{x})^2$$

Dans tous les cas : $\sigma(x)=\sqrt{V(x)}$

Droites de régression

$$\mathrm{Cov}(x,y)=\sigma_{xy}=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})=\frac{1}{n}\sum_{i=1}^{n}(x_i y_i)-\bar{x}\,\bar{y}$$

$y=ax+b$, où $a=\dfrac{\mathrm{Cov}(x,y)}{V(x)}$; $x=a'y+b'$, où $a'=\dfrac{\mathrm{Cov}(x,y)}{V(y)}$

Coefficient de corrélation linéaire $\quad r=\dfrac{\mathrm{Cov}(x,y)}{\sigma_x\sigma_y}$

II. COMBINATOIRE - DÉNOMBREMENTS

$\mathrm{Card}(A\cup B)=\mathrm{Card}\,A+\mathrm{Card}\,B-\mathrm{Card}(A\cap B)$

$\mathrm{Card}(A\times B)=\mathrm{Card}\,A\times\mathrm{Card}\,B$

Soit E un ensemble de n éléments

Nombre de permutations de E : $n!=1\times2\times3\times\ldots\times n$; $0!=1$

Nombre d'arrangements de p éléments de E :

$$A_n^p=n(n-1)\ldots(n-p+1)$$

Nombre de sous-ensembles de p éléments de E :

$$C_n^p=\binom{n}{p}=\frac{A_n^p}{p!}=\frac{n(n-1)\ \ldots\ (n-p+1)}{p!}=\frac{n!}{p!\,(n-p)!}$$

$$C_n^p=C_n^{n-p} \quad ; \quad C_{n+1}^p=C_n^p+C_n^{p+1}$$

$$(a+b)^n=a^n+C_n^1a^{n-1}b+\ldots+C_n^ka^{n-k}b^k+\ldots+b^n$$

III. PROBABILITÉS

Si A et B sont incompatibles : $P(A\cup B)=P(A)+P(B)$

Dans le cas général : $P(A\cup B)=P(A)+P(B)-P(A\cap B)$

$P(\bar{A})=1-P(A)$; $P(\Omega)=1$; $P(\varnothing)=0$

Si $A_1,\ldots,A_n$ forment une partition de A, $\quad P(A)=\sum_{i=1}^{n}P(A_i)$

Dans le cas équiprobable : $P(A)=\dfrac{\mathrm{Card}\,A}{\mathrm{Card}\,\Omega}$

Probabilité conditionnelle de A sachant que B est réalisé

$P(A\cap B)=P(A\,|\,B)\,P(B)$; $P(A\,|\,B)$ se note aussi $P_B(A)$

Cas où A et B sont indépendants : $P(A\cap B)=P(A)\,P(B)$

Formule des probabilités totales

Si les événements $B_1, B_2, \ldots, B_n$ forment une partition de Ω,

alors $P(A)=P(A\cap B_1)+P(A\cap B_2)+\ldots+P(A\cap B_n)$

Variable aléatoire

Fonction de répartition : $F(x)=P(X\leqslant x)$

Espérance mathématique : $E(X)=\sum_{i=1}^{n}p_i x_i$

Variance : $V(X)=\sum_{i=1}^{n}p_i\left(x_i-E(X)\right)^2=\sum_{i=1}^{n}p_i x_i^2-\left(E(X)\right)^2$

Écart type : $\sigma_x=\sqrt{V(X)}$

Loi binomiale : $X(\Omega)=\{0,1,\ldots,n\}$; $P(X=k)=C_n^k p^k(1-p)^{n-k}$

$E(X)=np$; $V(X)=np(1-p)$

Loi de Poisson : $X(\Omega)=\mathbb{N}$; $P(X=k)=\dfrac{\lambda^k e^{-\lambda}}{k!}$; $E(X)=\lambda$; $V(X)=\lambda$

Si X_1 suit la loi $\mathcal{N}(m_1,\sigma_1)$, X_2 suit la loi $\mathcal{N}(m_2,\sigma_2)$, et sont indépendantes, alors X_1+X_2 suit la loi $\mathcal{N}(m_1+m_2,\sqrt{\sigma_1^2+\sigma_2^2})$

Échantillonnage

◆ La variable aléatoire $\bar{X}$ qui, à tout échantillon aléatoire **non exhaustif** de taille n prélevé dans une population de moyenne m et d'écart type σ, associe la moyenne de cet échantillon, pour n assez grand, suit approximativement la loi $\mathcal{N}\left(m,\dfrac{\sigma}{\sqrt{n}}\right)$.

◆ La variable aléatoire F qui, à tout échantillon aléatoire **non exhaustif** de taille n prélevé dans une population présentant un pourcentage p d'individus possédant une certaine propriété, associe le pourcentage des éléments de cet échantillon qui possèdent cette propriété, pour n assez grand,

suit approximativement la loi $\mathcal{N}\left(p,\sqrt{\dfrac{p(1-p)}{n}}\right)$.

◆ Dans le cas d'échantillons **exhaustifs** les écarts types de $\bar{X}$ et de F sont multipliés par le facteur d'exhaustivité $\sqrt{\dfrac{N-n}{N-1}}$.

IV. STATISTIQUE INFÉRENTIELLE

Estimation ponctuelle

◆ Une estimation de la moyenne m d'une population, est donnée par la moyenne $\bar{x}$ d'un échantillon aléatoire non exhaustif prélevé dans cette population.

◆ Une estimation du pourcentage p d'individus possédant une certaine propriété dans une population, est donnée par le pourcentage f d'individus possédant cette propriété dans un échantillon aléatoire non exhaustif prélevé dans cette population.

◆ Une estimation de l'écart type σ d'une population, est donnée par $\sqrt{\dfrac{n}{n-1}}\,s$, où n est la taille et s l'écart type d'un échantillon aléatoire non exhaustif prélevé dans cette population.

Estimation par intervalle de confiance

◆ Un intervalle de confiance de la moyenne m d'une population d'écart type σ, avec le coefficient de confiance $2\Pi(t)-1$, est

$\left[\bar{x}-t\dfrac{\sigma}{\sqrt{n}}\ ,\ \bar{x}+t\dfrac{\sigma}{\sqrt{n}}\right]$, où n est la taille et $\bar{x}$ la moyenne d'un échantillon aléatoire non exhaustif prélevé dans cette population.

◆ Un intervalle de confiance d'un pourcentage p d'individus possédant une certaine propriété dans une population, avec le coefficient de confiance $2\Pi(t)-1$, est

$$\left[f-t\sqrt{\frac{f(1-f)}{n-1}}\ ,\ f+t\sqrt{\frac{f(1-f)}{n-1}}\right]$$

où n est la taille d'un échantillon aléatoire non exhaustif prélevé dans la population, et f le pourcentage d'individus possédant cette propriété dans cet échantillon.

◆ Dans le cas d'échantillons **exhaustifs** les intervalles de confiances décrits ci dessus sont respectivement :

$$\left[\bar{x}-t\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}\ ,\ \bar{x}+t\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}\right]$$

$$\left[f-t\sqrt{\frac{f(1-f)}{n-1}}\sqrt{\frac{N-n}{N-1}}\ ,\ f+t\sqrt{\frac{f(1-f)}{n-1}}\sqrt{\frac{N-n}{N-1}}\right]$$

Tests d'hypothèse

◆ Choix : - d'une hypothèse nulle H_0 ($m=m_0$ ou $p=p_0$)

- d'une alternative H_1 ($m\neq m_0$ ou $p\neq p_0$) **[test bilatéral]**

($m>m_0$, $m<m_0$, $p>p_0$, $p<p_0$) **[test unilatéral]**

◆ Recherche de la région d'acceptation ou de la région critique sous H_0 au seuil de risque α

◆ Énoncé de la règle de décision

◆ Application du test avec échantillon(s)

Auteurs *Bernard BIGOT, Brigitte CHAPUT,*
Jean-Claude DUPERRET, Jean-Claude DANIEL

Titre *Probabilités et statistiques - Statistiques inférentielles (B.T.S.)*

Editeur *IREM de REIMS*

Date *Octobre 1996*

Niveau *Lycée : sections de B.T.S.*

Mots-clés ~~*Probabilités, statistiques, statistiques inférentielles.*~~ à remplacer

Résumé *Cette brochure a été réalisée à partir des documents élaborés pour l'animation de stages de formation des enseignants de mathématiques des sections de Techniciens Supérieurs. Elle est avant tout destinée aux enseignants. La première partie pose le problème de la modélisation (passage des statistiques aux probabilités), la seconde aborde la notion de loi de probabilité et débouche sur l'échantillonnage en vue des statistiques inférentielles traitées dans la troisième partie.*

Brouillon

* Elle est le fruit des recherches du groupe BTS de l'IREM de REIMS et

TITRE : PROBABILITES ET STATISTIQUES - STATISTIQUES INFERENTIELLES (B.T.S.)

AUTEUR : B. BIGOT, B. CHAPUT, J.C. DUPERRET, J.C. DANIEL

NIVEAU : Lycée : sections de B.T.S.

EDITEUR : IREM de Reims

DATE : Octobre 1996

MOTS-CLE : spécialité PROBABILITES, STATISTIQUES, STATISTIQUES INFERENTIELLES

autres
- Estimation ponctuelle et par intervalle
- Loi (de probabilité)
- Statistiques et statistiques inférentielles
- Test d'hypothèses
- Variable aléatoire

RESUME : Cette brochure a été réalisée à partir des documents élaborés pour l'animation de stages de formation des enseignants de mathématiques des sections de Techniciens Supérieurs.

Elle est le fruit des recherches du groupe B.T.S. de l'IREM de Reims et est avant tout destinée aux enseignants.

La première partie pose le problème de la modélisation (passage des statistiques aux probabilités), la seconde aborde la notion de loi de probabilité et débouche sur l'échantillonnage en vue des statistiques inférentielles traitées dans la troisième partie.

ISBN 2-910076-10-5

FORMAT	NOMBRE DE PAGES	PRIX	IREM numéro
A4	174	75 F HT + 16 F (frais d'envoi)	Re37